2019

Talking Red White and Blue: An Investigation into the Relationship Between Polarization and Congressional Floor Speech

Dawson Patrick Honey
The College of Wooster, dhoney19@wooster.edu

Talking Red, White, and Blue

An Investigation into the Relationship Between Polarization and Congressional Floor Speech

By: Dawson Honey

An Independent Study Thesis
Department of Political Science

The College of Wooster
March 2019

Submitted in partial fulfillment of Senior I.S. Thesis

Advisor: Dr. Desiree Weber
Second Reader: Dr. Bas van Doorn

TABLE OF CONTENTS

Table of Figures

Acknowledgements

This project required me to learn many new skills that I will carry with me for the rest of my life. There were many times during this project where I would have crashed and burned had it not been for the help of peers, professors, and friends. Before proceeding I would like to thank all those who assisted me and helped my project succeed.

I would first like to thank Dr. Marian Frazier. Her help was essential in figuring out the coding for this project. When I began this project, I had only a beginner's understanding of R Studio, the software used for all of my statistical analysis. Coincidently, the small amount of R coding I did know I learned in Dr. Frazier's basic statistics class during my junior year. When it came time to start my Senior IS, I knew that I wanted to use R Studio since it fit best for the construction of my project. Many of the methods of this thesis required techniques and knowledge which was far beyond my current understanding of the software. Dr. Frazier was able to help me construct the statistical elements of this paper that allowed me to construct my project without any concessions; Without her advice my project would be very different.

I would next like to thank my good friends Maxx Coburn and Kevin Thomas. Just as I was starting to gather data for this thesis, the database I was using for congressional speeches was retired. This made the software that I was using inoperable, as it was not designed to work with the new database. I did not possess the skill or the time to figure out how to modify the software since it was in a programming language with which I had little experience. Maxx and Kevin are seasoned coders and agreed to help me find a solution that would allow me to complete my project.

Next, I would like to thank Nicholas Judd, the co-creator of the Congressional Record Parser that I used to gather all my data. At the time of this project, he was in charge of maintaining the script and making sure that it remained in working order. When the database changed, his prompt help and advice was a vital in making the script work with the new congressional database.

I would like to thank my advisor, Dr. Desiree Weber. Even though my topic was not in her primary realm of expertise she still provided me with ample guidance and advice about my project. Early in the year, she showed me with an article that where I discovered the existence of the python script that helped me gather data.

Lastly, I would like to thank Dr. Bas Van Doorn, my second reader. His help during my junior IS set me on the right course for my senior year. I was able to begin my IS with much of the groundwork already laid out, which made the process less strenuous.

Introduction

A word is not a crystal, transparent and unchanged; it is the skin of a living thought and may vary greatly in color and content according to the circumstances and time in which it is used.

-Justice Oliver Wendell Holmes, Jr.

For the latter half of the 20th century, the estate tax was the most basic form of a common-sense law. It was a tax on the multi-million-dollar inheritance of our nation's most wealthy citizens which would seldom be a burden to the everyday voter. On Capitol Hill, the prospect of repealing the estate tax was a forlorn goal kept alive only in the minds of the most ambitious fiscal conservatives. After all, it is very difficult to convince the voting populace that repealing a tax on wealth inheritance is in their best interest. But starting in the late 1990s, a coalition of legislators and activists began a coordinated effort to turn the public against the tax. One of their most effective tactics was to reframe the issue in a way that portrayed the tax as cruel and ambivalent to human suffering. This effort was crowned by renaming the estate tax to the pithy moniker: "death tax." Combined with aggressive speechmaking and fervent public outreach, conservative representatives began to speak out in opposition of the estate tax. In response to these campaigns to repeal the death tax, masses of small business owners, farmers, and day laborers demanded that Congress allow large fortunes to be passed on with minimal taxation. Those who called for the repeal most fervently were not affected by the tax in the slightest, but they still demanded its demise as if the survival of their livelihood hung on its abolition. (Graetz and Shapiro 2005, 3)

While coining the term “death tax” alone cannot be credited with turning opinion against the tax, it demonstrates the power of word choice in modern political discourse. With only a few soundbites, well placed advertisements, and incessant repetition, congressional policy can be turned against even the most bland and mundane policies. Talking points that may initially come across as nonsense will begin to seem more sensible with proper phrasing. The chambers of Congress are a potent source for persuasive rhetoric that emanates outward into the public discourse. Congressional floor time provides an opportunity to transmit party or personal messages. Floor time has utility both as a method to persuade colleagues on either side of the aisle and as a way to promote a representative’s image to constituents.

The question of this research project relates to the relationship between word choice within one-minute floor speeches and political polarization of Republicans and Democrats. The goal is to determine whether increases in political polarization and interparty unity correspond with more consistent word choice among party members. Polarization, or the ideological division between political parties, has become an increasingly prominent topic of discussion in modern politics. The debate surrounding polarization creates divisions among voters and cuts to the core ideological predications of our political discourse. Congressional polarization has established itself as one of the most enigmatic obstacles to political progress. In the eyes of many, it is the harbinger of gridlock and political stalemate where the dynamics of party politics are king.

Polarization is not an easy issue to address because doing so forces one to confront basic assumptions about how politicians and parties operate. For instance, a key concern of rising polarization is the ultimate demise of friendly discourse in favor of recalcitrant, unrelenting, stubbornness. However, at the same time, many voters say they dislike compromise. In a 2018 study, Pew Research found that roughly half of Americans (53%) would prefer their political representatives to stick to their positions rather than make a compromise with the other side. This study also showed a substantial partisan divide on this issue. Whereas Republicans have remained around 30-40 percent in favor of compromise, the Democrats consistently are more supportive of comprise, with a peak of almost 70 percent in 2017. However even the Democrats have dropped below 50 percent since 2017, indicating a souring attitude towards working with Republicans. (Pew Research Center 2018)

A well-recorded symptom of polarization is not just the growing distance between parties, but also growing ideological consistency within each party. As Democrats and Republicans get farther apart from one another, each party's members get more similar to each other. The question then becomes: in an environment where individuals are more similar in their beliefs, will this correlate with similar speech-patterns and word choices? This analysis treats each party as an independent case study, and rather than looking at their interactions mainly focuses in the word choice differences that flow from a more homogenous ideological environment. The hypothesis of this study is that when political polarization is higher, this will correspond with more similar speechmaking among representatives. Specifically, the analysis of this study examines the word frequency in floor speeches. The word frequency of a body of speeches refers to the concentration of common words. Word frequency is calculated by finding the distribution of the most common words as a percentage of the whole body of speeches. In order to create a consistent data frame for accurate comparison, the data will consist of each speech on the floor of the House of Representatives since 1995 divided into six-month data chunks.

Chapter one of this thesis establishes the main currents in the literature that have covered polarization and word choice. In the realm of polarization there has been much study on how to measure polarization, with many sources using the standard roll call voting scores as a metric of ideological distance. However, for this study, there will be an additional metric of polarization. This metric uses coverage of polarization and gridlock in a conglomerated list of news-media sources to gauge the prominence of polarization in the public. The use of roll call votes provides a metric of polarization that focuses on the conduct of the legislators while the second metric of media coverage gives an outside-in approach relating to the perception of polarization. Both of these metrics will be used to identify eras of high and low polarization that will be compared to the distribution of words over time.

Also discussed will be the history and causes of polarization in order to provide context and perspective. While not a functioning element in the variables of this study, the genesis and progression of the ideological landscape of Congress provides vital information to evaluate the results of this study. The two key observations about the history and causes of polarization are, first, that polarization is a longstanding trend that is not a recent phenomenon. Most scholarship supposes that the modern trend of rising polarization has been present since at least the 1950s in the aftermath of World War II. It is not just an issue that besets our most recent Congresses but has been a strong player in political discourse for many decades. Secondly, this chapter will also argue that polarization can be present with or without gridlock and dysfunction. There is nothing about the nature of polarization that requires that both parties to have equal voting power. If one party has a substantial majority, and uses it to great effect, then bills will still be moving through Congress even in a highly polarized environment. This was the case for much of the postwar period, where Republicans experienced one of the longest eras of minority power in their history.

Another prominent discussion in congressional polarization is its reciprocity with the public. In other words, how much of the polarization of Congress is reflected in the minds of the average American, or vice versa. The question here is whether polarization represents a shift in the general public's perception of government or if the ideological divisiveness of Congress ends at the Capitol steps. The topic of whether the public is actually polarized has been a prominent topic in political science for decades and the literature on this topic is extremely divided. Many studies find that the public is just as polarized as their elected officials while others find the public to be indifferent to the ideological dissimilarity of their representatives. Measuring the polarization of the public is problematic because in order to do so you must be clear what parts of the nation count as the public. For example, a relevant question is whether ardent partisans and political activists are considered in the analysis of the public or if the evaluation is limited to just those whose political participation is limited to the ballot box. Amongst dedicated partisans, there has been a measured increase in ideological distance between the two parties that reflects the divisions of Congress. Still, these individuals represent a minority of the public at large and should not be purported to represent public opinion.

For most of the studies that found polarization in partisan actors, the ideological distance is quantified using deviance or conformity to values which are established by each respective party platform. When dealing with the public at large, measuring ideology becomes more troublesome due to the fact that they have no roll call votes to use and generally are not acting with a party platform in mind. The problem then becomes one of deciding which issues and concepts will serve as the fulcrum to measure ideological distance. The public has a diverse collection of opinions that are all motivated by different reasons. In such a vast expanse of different views justified by a myriad of experiences, it may be impossible to truly evaluate where the country wide ideological center is and who falls on which side.

It is not the intention of this paper to make a definitive argument on the issue of public polarization because doing so would not be have a strong enough impact on the outcome on the results of this analysis. This study is inherently a study of the dynamics and behavior of Congress, with the behavior of the public at large set as a backdrop. But the question of whether the public mirrors Congress is still relevant in evaluating the kinds of speeches that representatives choose to make. The power of speechmaking and the modern availability of information creates an environment where members of Congress use floor time to convince their constituencies just as much as their colleagues. The presence of constituent communication is a relevant factor in this research since it may be affected differently by the paradigm of party politics.

Chapter two describes the methods and coding of the study. The primary data sources for measuring polarization come from roll call voting and media perception of Congress. Together these metrics provide a trend line that charts the amount of polarization in Congress. Against this trend line of polarization, this study compared the trends in word frequency to assess the degree of convergence between polarization and the distribution of common words. For speech data, records will be collected from the *Congressional Record* and analyzed using statistical analysis software. Speech data is available all the way to 1995, so this analysis focuses on Congressional speechmaking since the 104th Congress. By far the most labor-intensive part of this analysis was gathering all the speech data from the *Congressional Record.* To accomplish this task, I used a Python script that has been specifically designed to scrape data from the Congressional database and parse it out into machine readable data. The script collects data in a chronological manner where all speech data from the one-minute House speeches is collected and parsed between two selected dates. For this project, each six-month chunk of speech data will result in one data point. The usage of this Python script allows every single speech since 1995 to be included in the analysis. However, the speeches collected are limited to one-minute floor speeches in the House of Representatives to ensure that long speeches or filibusters did not induce any skewing in the speech data.

This study is primarily a correlation study where the trends of polarization since the mid-1990s will be compared to the trends in speech similarity within the two major parties. Speech similarity, or, word frequency, is defined as the distribution of the most common words on a descending order scale. The most common words in a data set will be at one end and the least common at the other. If a larger percentage of words are clustered towards the more common end of the chart, that indicates higher speech similarity within the party. This is the case because the common words make up a larger portion of the whole body of text.

For this study there are two different metrics of word frequency. The first measure uses the top one hundred most commonly used words for each party. The top words are added up and divided by the total count of words in the whole speech database to create a percentage. The percentage that the top one hundred words represent of the whole creates that time span's frequency score. The second method of quantifying word frequency uses a small collection of specific, politically charged words, to capture partisan speechmaking in a way that uses a consistent set of words. The words chosen for this method are ones that are commonly used and reflect each party's platform and common talking points. These "meat words" were divided into positive and negative words, and separate analysis was conducted on each. Calculating the frequency score is the same as the top one hundred words metric: the counts of the selected words are added up and placed as a percentage of the whole. This percentage frequency score is calculated for each six-month set of speech data in order to establish a trend line for word frequency.

Chapter three describes the results of the three metrics of word frequency for each party and discusses how they fit with the models of political polarization. Also discussed is how the word frequency scores compare to eras of party control and the instances of presidential elections. This is done to explore additional correlations in word frequency besides polarization. In the data frame of this study, there have been three distinct points where party dynamics shifted. The first point was in 1995 when the Republicans regained a majority in the House for the first time in decades. The second point started in 2006, when the Democrats regained their majority after gaining around thirty seats. The final point begins with the 2010 election when the Republicans once again regained majority voting power in the House.

Chapter four is a discussion of the results of this analysis to evaluate whether there is enough evidence to justify a positive relationship between polarization and the distribution of words in Congressional speechmaking. The finding is that polarization and word frequency are not strongly correlated. Instead, the data demonstrate that speechmaking is much more strongly affected by which party is in control and presidential elections. For the Republican party, the distribution of words across both metrics was strongly correlated with losing and gaining back control of the House. For Democrats, their word distribution was more strongly correlated with whether their party controlled the White House. This section also speculates about how the techniques and procedures of this study suggests avenues for future research. Perhaps the most substantial contributions of this analysis are the methods it establishes for evaluating word distribution over time: This construction could provide a platform to evaluate a host of other political questions related to word choice and rhetorical consistency.

Chapter 1: Literature Review

Defining Polarization

Before any extensive analysis can be done, it is important to define what is meant by the term "polarization." The term at its most basic level refers to the ideological division between the two major parties in Congress. In a more polarized Congress, each party will consist of fewer and fewer moderates. Polarization can be illustrated with a bimodal distribution, or a double-humped bell curve. If a standard bell curve represented an ideological breakdown where the majority of legislators are close to the center, polarization occurs when the two sides pull apart, resulting in fewer moderates and a large concentration of representatives on either side of the median. The core and leadership of each party moves away from the center, with each side becoming more radical. In short, polarization refers to the separation of Democrats and Republicans into liberal and conservative camps. In addition to moving farther from each other, each party has seen a decline in moderation. Each party has become more ideologically clustered surrounding the more extreme ends of their respective ideological spectrums. (McCarthy, Poole, and Rosenthal 2016, 4)

Polarization is a two-faceted issue. On one hand there is the difference between the parties, and on the other, there is the difference within each party. As the parties move farther from one another, each party member gets closer to their peers. This tends to develop concurrently with the distance between parties and creates more ideologically consistent, separated parties. This increasing ideological consistency that persists in intra party relations is a key factor in this analysis. The main goal of this study is to examine how speechmaking is influenced by growing ideological distance between colleagues and similarity between peers.

The definition of polarization does not include any judgements about gridlock or lack of progress on key issues. Polarization merely refers to the ideological gulf between parties, which need not necessarily result in a stalemate. If a party retains majority control, then it has more ability to move legislation. This in turn, can minimize the influence of a politically polarized Congress. The current pattern of partisan control is characterized by a Congress that is frequently shifting back and forth from Democratic to Republican hands with both parties in roughly equal numbers from year to year. This constant switching of power and competition of ideas has thrust discussions of the effects of polarization into the forefront of political scholarship.

Measuring Polarization

Perhaps the most direct way to measure polarization in Congress is through the use of roll call voting. Roll call voting is a common method for evaluating the ideological expressions of Congressional Representatives. This method is named the Dynamic Weighted Nominal Three-step Estimation (DW-NOMINATE) and uses the records from roll call voting to create a scale for representatives' ideological positions on a liberal to conservative scale. The process measures ideological positions on two issue areas: social and economic. DW-NOMINATE scores are a common means of scholarly evaluation of polarization, partly because the records provide a massive database that allows for comprehensive models of the position of each legislator that allow for repeatable, reliable statistical analysis. (Poole and Rosenthal 2018) Political science scholarship has been using these scores to conduct analysis of Congress for over twenty years. (Carroll et al. 2009) DW-NOMINATE scores provide a convenient base point to analyze polarization. Since roll call voting is well documented, it creates the potential to track polarization for the entire history of Congress. Using DW-NOMINATE scores will provide a reliable way of measuring polarization over time.

While DW-NOMINATE provides a way to numerically track the voting behavior of Congress and infer the ideology of the representatives, that does not necessarily capture the full breadth of polarization. For this study, these values are accurate, but not necessarily reliable. Many scholars disagree about how to measure polarization and from what angle. Some say that polarization is best measured in a Congress to Public relationship, where polarization in Congress in turn causes polarization of the public. Others see it as the opposite, where influences of a divided public will cause Congress to become more polarized. DW-NOMINATE scores only capture the former dimension. For the purpose of this research, it is best to not rely too heavily on one interpretation of polarization, rather, it is better to use multiple different analyses and then confer between them to create the best picture of when polarization is the most prevalent.

The second way that polarization will be measured for this study relates to the news media's perception of how polarized Congress is. In her 2013 paper, Maria Azzimonti constructed the Political Polarization Index (PPI) by measuring the frequency of newspaper coverage relating to political disagreement. The more news outlets included words that related to gridlock or divided government, the higher that year scored on the index. According to this measure, polarization has been increasing since the 1990s with the most recent data being the highest measure of polarization in the 60 year span of the study. This trend is shown in *Figure 1*. (Azzimonti 2013, 5)

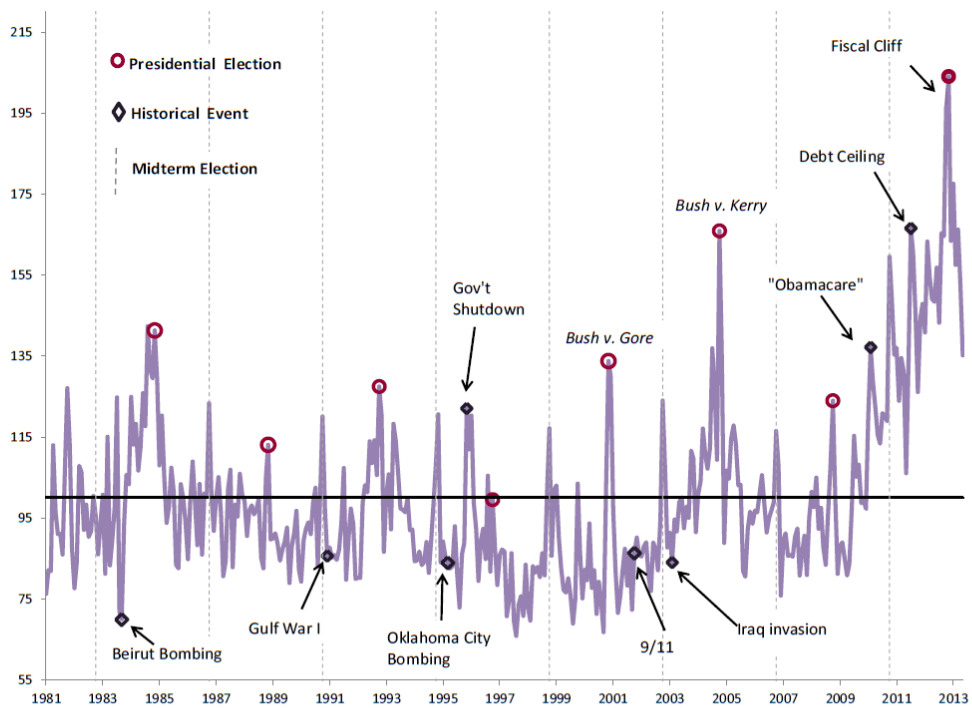

Figure 1: Political Polarization Index (PPI) (Azzimonti 2013, 5)

Polarization affects legislation in many significant ways, both externally and internally. Another means of studying the degree of polarization is to study the amount of disagreement between parties and the instances of gridlock. David Jones studied the polarization of representatives using voting history and deviations from party line votes. He found that higher levels of polarization correspond with higher levels of legislative gridlock. (Jones 2001, 136) However despite the fact that there was this trend, the presence of polarization alone does not mean that gridlock is inevitable. Gridlock requires a roughly even split between both parties, but such an even split is not intrinsic to polarization. It is entirely possible to have a polarized Congress where one party still holds a substantial majority. (Jones 2001, 128)

This points out a significant assumption in polarization research: that gridlock and polarization are the same phenomenon. While Jones shows that both do often arise concurrently, polarization does not by itself create gridlock. In formal logic terms, while polarization may be one necessary element for gridlock, it is not sufficient on its own. Despite this, polarization and gridlock will go hand in hand more often than not. Jones also notes that since 1990, more than half of every congressional vote has consisted of a majority of one party, indicating that party members are more often voting as single bloc with little dissent on either side. (Jones 2001, 125)

Jones' article alludes to the "divided government hypothesis," which may have implications for this study. The hypothesis states that when a president's party does not control the House or the Senate, legislation is less likely to be enacted. This may have an impact on this analysis since a House that is at odds with the President may result in more consistent word choice. A divided Congress where the chambers of Congress are in competition with the executive branch will magnify the influence of contentious party issues and may create more unity among party members. However, while intuitive, this hypothesis relies on three key assumptions. First, that passage of legislation always requires a simple majority. Second, that it is impossible to pass legislation unless the President and Congress agree. Third, that each party will always have diametrically opposed preferences. (Jones 2001, 126) While these assumptions may ring true for headline politics, studies on the effects of a divided Congress produce mixed results. Systematic analysis of significant laws in the postwar period found a divided congress does not reduce significant legislation. (Mayhew 1991) Other studies using similar data have found that divided government results in less legislation when accounting for non-stationary time series data. (Kelly 1993) Further research has even suggested that although others have found that while it reduces landmark legislation, it actually increases the passage of less significant legislation. (Cameron et al. 1997)

This study is primarily concerned with the usage of words in Congressional speeches and how that usage relates to the level of polarization. A logical next step would be to examine how polarization can be quantified from a word-choice perspective. Monroe et al provide salient research on partisan word choice by using statistical analysis to track the usage of particular politically charged words. The analysis demonstrated by *Figure 2* shows that certain words vary in their usage by either party. This research provides a measure of polarization that is represented by the polarization of specific words. The study uses the difference between the amount a particular word is used by each party to determine which words have become more associated with one party and which ones are used by both. If a set of words is used roughly equally by both parties, it indicates a minimal amount of polarization. However, if certain key words are more associated with one party, that corresponds with higher polarization. An exemplary case study is the word "Iraq." From 1997 through 2001 the word remained largely neutral. However, after 9/11, the word became used much more often by Republicans than Democrats. After the War Authorization and invasion of Iraq in 2003, the word became strongly charged for Democrats. As illustrated in *Figure 2*, plotting multiple words on a single chart reveals a more cogent measure of polarization. Words like "budget," "defense," and "education" swing back and forth from being used more by one party to being used in largely the same amount by both.

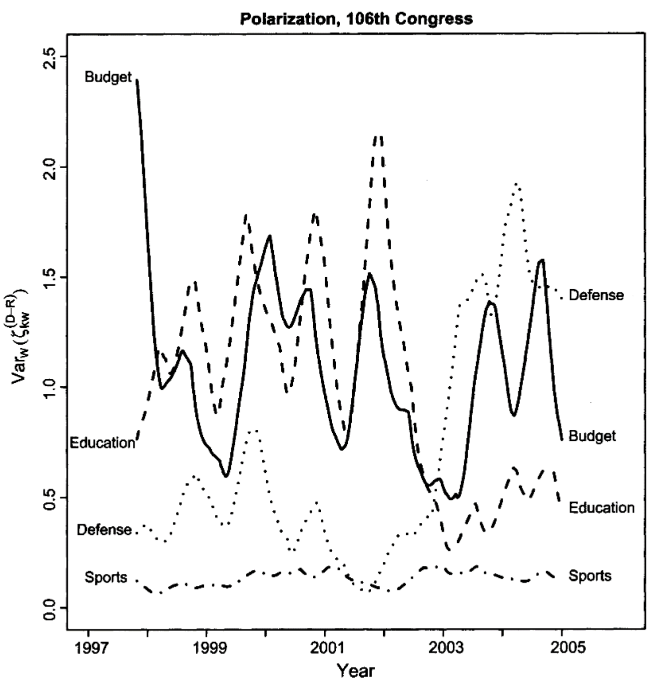

Figure 2: Variance in party word association 1997-2005 (Monroe et al 2008, 398)

Interestingly, but perhaps not surprisingly, the word "sports" remains level throughout the study, with very little difference between the amount of usages by each party. (Monroe, Colaresi, and Quinn 2008, 398) The study only spans from 1997 to 2005, so its scope is limited. However, it does present evidence in favor of the hypothesis of this study as it shows that political conditions such as polarization do correlate with the consistency of Congressional word choice. When higher polarization is present, that corresponds with one party using certain words more than the other. If this trend holds for word-choice in general, then it follows that highly polarized partisan eras will correlate to more similar speechmaking by members of a particular party.

For this study, the sampling will occur from 1995-2018. According to the literature, the 23-year period will encompass some of the most substantial and impactful eras of polarization. Rather than picking just two points in time to compare, it will utilize all the speeches that were delivered on the floor between those two dates to create a large corpus of congressional speeches. To operationalize polarization for this analysis, the two metrics are DW-NOMINATE and the PPI. To operationalize word choice, this analysis will create two separate metrics using the top one hundred most frequent words and the frequency of specific partisan words over time.

Causes of Polarization

The causes of political polarization provide a useful reference point to evaluate the effects of polarization. Thus, much study has been devoted to its origins. Scholars point to a variety of reasons behind the rising trends of low party cooperation. One such reason is the tactics and strategy of Newt Gingrich and the Republican Party during the late 1980s to early 1990s which incentivized legislators to fall in line with party orthodoxy. Another, less specific cause was the Republican and Democratic replacement and realignment of old policy combined with the influx of new legislators who were more in line with party identity. (Roberts and Smith 2003, 315)

Almost as important as the causes of polarization is what did not cause it. For instance, it might stand to reason that partisan gerrymandering played a role, since it allows for the creation of safer districts and less representative accountability. However, gerrymandering only works for offices that use districting; it does not explain how the Senate became polarized. The Senate shows similar or more extreme levels of polarization despite the fact that gerrymandering has no effect on its elections. Though increased gerrymandering may be an effect of increased polarization, it cannot explain the full breadth of polarization in government. (Campbell 2016, 147) Another commonly attributed cause of polarization is the acerbic partisan media. When it comes to political issues in the United States, "The Media" is a common pejorative catch-all explanation for any and all societal complaints. At first glance, the arguments in favor of media influencing polarization are not without merit. Influx of media sources, some argue, prevent politically interested citizens from gaining meaningful access to politically discordant sources. The polarization of media sources will then result in two distinct messages filtering through to the American public who then in turn elect more polarizing representatives. (Campbell 2016, 150). However, the youth of polarized media insulates it from the eras of political time where polarization began. Rush Limbaugh and Fox News, for example, only began broadcasting in the late 80s and 90s, whereas the rise of contemporary political polarization can trace its roots all the way back to the 1950s.

Polarization in Congress

Polarization has been a dominant theme in political analysis of Congress for the last twenty years. A glance at the DW-NOMINATE scores for the 115th Congressional class from 2017 to 2019 shows a clear rift between Republican and Democratic representatives. There are no outliers who do not vote in line with the party mass. While some may be more extreme than the norm of their party, there are not any who cross the threshold to the other side of the ideological spectrum, as shown in *Figure 3*.

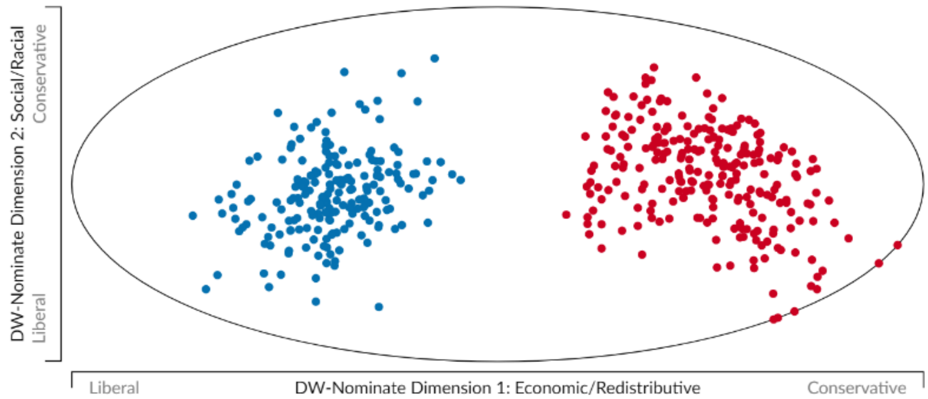

Figure 3: DW-NOMINATE for the 115th Congressional Class (voteview.com)

Looking at the voting records for the 95th Congress (1977-1979) demonstrated by *Figure 4* shows a much larger spread, both among the House of Representatives as a whole and among each party. In the 95th Congress, there were socially conservative Democrats and socially liberal Republicans. The same is true, although less so, for economics issues. Outliers existed in each party, in contrast to modern Congress, where the parties are two homogenous clusters within the confines of their respective tree houses.

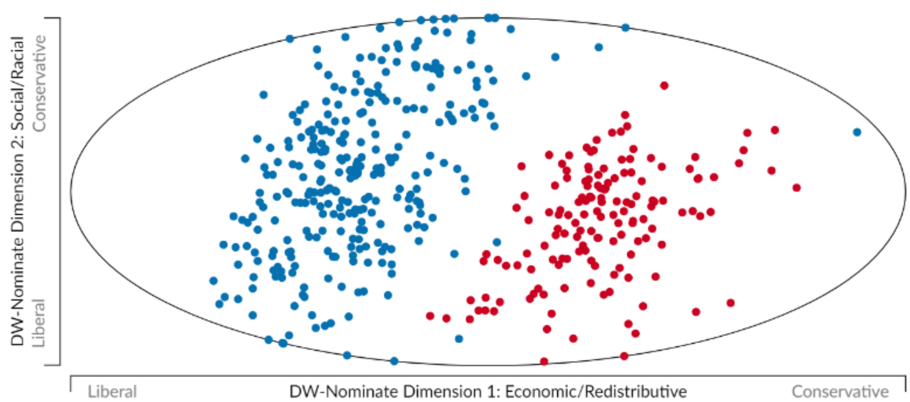

Figure 4: DW-NOMINATE for the 95th Congressional Class (voteview.com)

Polarization is not a one-dimensional issue: There are many issue areas that representatives may disagree on, some of which are more contentious than others. This presents a potential complication to measuring polarization. Economics and cultural issues are the two issue-area lenses of polarization, both expressed in subtly different ways. Concepts like the size of government and equal opportunity have remained relatively consistent with large gaps between the amounts of support from each party. However, concepts like taxes and free markets have seen increasing polarization since the 1970s, when the gap between the party members was relatively small. (Wood and Jordan 2017, 218–21) When it comes to social and cultural issues, there is also an increasing divide. A measure of partisan attitudes towards environmentalists shows that in the 1990s, around 76% of Democrats and 74% of Republicans had a favorable outlook. By 2010, those figures had split to around 55% and 71% respectively. The same divide is present in abortion rights, in the 1970s, both parties were around 60% pro-choice. Over the next 40 years, Republicans dropped to less than 50% while Democrats remained roughly the same. (Wood and Jordan 2017, 222–27)

Of the two issue areas, the most divisive issue of modern politics is economics. Since the 1980 election of Ronald Reagan it has been the dominant issue area that separates the parties. The conflict has escalated through issues over who should benefit from the government and how to pay for it. Since 1989, no other issue has divided partisans like the annual federal budget debates, Republicans favoring tax cuts and reduced expenditures, whereas Democrats uniting under a platform of deficit reduction though increased taxes on the upper class. (Wood and Jordan 2017, 200) Logically, this intense divide makes sense. Contrasting with social issues, where a division might arise due to a disagreement about where and how to apply the founding concepts and liberties of our nation, disagreement about economics come from two contradicting conceptions about the role of government. Fiscal liberals might argue that it is the role of government is to provide social services, whereas conservatives would argue that the role of government is to encourage the private market to provide those services. These "hands-on" or "hands-off" approaches represent two diametrically opposed visions for government action.

Polarization in the Public

There is considerable scholarly debate about whether the polarization of Congress is reflective of the political attitudes of the common citizen. Additionally, there is disagreement about whether Congressional polarization can trace its origin to a more polarized public, or a divided public is caused by a divided Congress. The first political observer to view polarization though the lens of the public was a sociologist named James Davison Hunter. He argued that American political strife was driven by a deep-seated disagreement about the founding principles of our nation:

> [U]nderneath the myriad political controversies over so-called cultural issues, there were yet deeper crises over the very meaning and purpose of the core institutions of American civilization… debates concerning the wide range of social institutions amounted to a struggle over the meaning of America. (Hunter et al. 2006, 14)

Some argue that a polarized public is one of the main contributing factors to a polarized Congress. From the 1950s to the early 1960s, the public was far less polarized, due in part to the galvanizing legacy of the Great Depression and World War II. But the 1960s saw the rise of a political upheaval accompanied by a new generation of voters who lacked the unity which bound voters who grew up during the 1930s and 40s. The 1960s and 70s were defined by conflicts like the Vietnam War and Watergate which sparked intense debate and dissonance among the new class of voters. Those voters would go on to elect representatives who shared their sympathies, resulting in the parties becoming increasingly at odds. (Campbell 2016, 152) The conceptual framework provided by Campbell suggests that a good era to observe a Congress prior to the influx of polarization would be during the 1950s, since that was still while the voter base had collectively experiences unifying national crises. Conversely, a proper time to observe the effects of high polarization would be divisive national crises. Looking at reoccurring debates during times of divisive and unifying national conflicts provides a way to observe the effects of polarization on the everyday debates of Congress.

There is circumstantial evidence to support the idea that the public is increasingly polarized. American National Election Studies (ANES) surveys on partisan voters indicate that in 1972, around one fourth of Democrats self-identified as liberals. Over the span of around forty years, that number grew to 49%. On the Republican end, those who identified themselves as conservatives grew from 42% in 1979 to 74% in 2012. (Campbell 2016, 119) ANES data additionally shows a clear rise in partisan membership corresponding with a decrease in split ticket voting. These data show that over the past several decades, party membership has risen from 26% to 33% while the percentage of voters who split their votes between the two parties has fallen from 28% to 17%. While these figures do not represent the entire public, it provides some credibility to the claim that politically active citizens are falling into party lines. (Campbell 2016, 121)

ANES data is corroborated by studies from other research institutes. In a study of the American public, Pew Research found that since 1994 the number of independents and nonpartisan voters has decreased substantially. Their study found that twenty years ago, “23% of Republicans were more liberal, than the median Democrat; while 17% of Democrats were more conservative than the median Republican. Today, those numbers are just 4% and 5%, respectively.” (Pew Research Center 2014) A key question that arises from these observations is an issue of causality. The data do not explain whether a polarized public creates a polarized Congress, or vice-versa. Additionally, the two phenomena could potentially develop independent of each other, resulting from a third, unknown source.

In addition, research has also suggested that the attitudes of partisan actors and activists has become increasingly bitter and vociferous. In 2014, Pew Research Center conducted a study on Democratic and Republican voters and found that the two had grown farther and farther apart from one another since the mid-1990s. Additionally, around 30 percent of both parties now view their political counterparts as a threat to the well-being of the nation. (Pew Research Center 2014) While these results might seem to point to a strong polarization of the public, these studies were conducted on the ardent political followers of both parties who are more in tune with the platform of their respective groups. It is not surprising that such groups would more closely mirror the ideological divide of the politicians they bolster.

While polarization is clearly evident in government and those who closely follow it, the attitudes and beliefs of the public at large still remain elusive. The polarization of enthusiastic political supporters should not be taken to represent the average citizen. In fact, it is entirely plausible that the populace at large takes no strong stance in partisan bickering. Following this line of thinking, some claim that the polarized public is a mirage of hyperbole brought on by tunnel vision on only the most extreme representatives of each ideology. Fiorina et al state that "[m]any of the activists do, in fact, hate each other and regard themselves as combatants in a war. But their hatreds and battles are not shared by the great mass of the American people." (Fiorina, Abrams, and Pope 2011, 8) However, a studied trend in the public is that voting districts across the country are becoming increasingly homogenous. During the 1976 election between Jimmy Carter and Gerald Ford, around 25 percent of the nation lived in a county where either candidate received more than 60% of the vote. By contrast, in the 2004 election between George Bush and John Kerry, almost 50% of the country lived in a county where either candidate received more than 60% of the vote. In around thirty years, nearly half of voters lived in a county with a disproportional percentage of Democrats or Republicans. (Theriault 2008, 3) The issue of whether the public is polarized factors into the decision regarding what eras of polarization to sample. Rather than taking a particular stance on the issue, what is most important is to create a sampling method that derives from the confluence of both perspectives. The eras of politics where levels of supposed public polarization line up with polarization in Congress provide cogent samples of polarized rhetoric.

Party Loyalty

This study is primarily a study of the relationship that group dynamics and ideological similarity have on Congressional word choice. Party loyalty is another lens to explore the divisions in Congress and its potential influence on word choice. A more polarized Congress will create two diametrically opposed parties, each becoming more uniform in their beliefs. It follows that party unity and loyalty may also factors into word choice. Those with similar views are more likely to espouse similar proposals; a more homogenous party could create an exchange of words and phrases between like minded Representatives.

A logical barometer for the loyalty among party members is roll call voting. When the party members think more alike to one another there will be less deviation from party-line voting. Much study has been devoted to measuring roll call loyalty. *Congressional Quarterly* publishes yearly reports on the unity of party members, using their roll call votes in favor or against their own party to calculate a unity score. Reviewing voting loyalty from 1987 to 2013, the party unity has increased substantially. In 1989, Republicans and Democrats voted for their own party's bills only 72-79% of the time. By comparison, the Congressional class of 2009 has party unity scores ranging from 91-92%. Additionally, as party unity increases, the standard deviation of party voting decreases, with members becoming increasingly clustered at the loyal end of the spectrum. (Box-Steffensmeier and Canon 2015, 54) Party unity provides another dimension to examine polarization, examining the word choice of congressional speeches during eras of higher party unity allows for analysis that explores the relationship between more consistent voting and consistency of word choice.

Another dimension of party unity is rhetorical party cohesion. Floor speeches are a means to bolster the party platform while also tearing down the opposing party. Contemporary scholarship suggests that one-minute floor speeches are as much an expression of party as an expression of the individual legislator. Unlike their colleagues in the Senate, House Representatives are often limited in the scope of their speech, both in content and in time. The only time when Representatives are free to speak on a subject of their choosing is at the beginning of each day during one-minute floor speeches. (Box-Steffensmeier and Canon 2015, 60) In recent years, one-minute floor speeches have become a vessel for coordinated partisan attacks. Party members, especially in the minority party, are far more likely to use their allotted floor time as a vantage point to attack the opposite party then praise their own.

When considering how to spread their message through one-minute speeches, both parties utilize inter-party organizations which coordinate party messages during floor time. These organizations, the Democratic Message Board and the Republican Theme Team, make requests for certain representatives to make speeches as well as directing them about speech subject matters. (Harris 2005, 127) When a representative is giving a one-minute speech, it is not only an expression of their personal interests but also reflects the party platform. These Party Message Organizations have a fairly recent history. They were both established between the one-hundred and first and one-hundred and second Congressional classes. Since their creation in the late 1980s to early 1990s, they have been working to stress repetition of key party phrases and positions in an effort to persuade representatives to participate and stick to the party message. (Harris 2013, 98) Much of their rhetoric has had the average American as their intended audience. Their message is not intended to persuade the other party, but rather is meant to communicate with middle America via C-SPAN. The stated intent of the Republican Theme Team is to "present the American people with a unified message on certain Republican themes." (Harris 2013, 99) The research on Party Message Boards confers with research done since the introduction of C-SPAN that articulate that the intended purpose of floor speeches has shifted in recent Congressional Classes. Before the introduction of C-SPAN and other services which broadcast Congressional proceedings to the public, the focus of much floor speech was to communicate with fellow representatives. After the introduction of C-SPAN, the purpose of floor speeches shifted to favor constituent communications. (Maltzman and Sigelman 1996, 820)

Word Choice

Rhetoric is how legislators communicate with and convince the public or their peers. Studying the consistency of rhetoric speaks to the ideological motivations that underscore the articulation of a political position. The question of this research is about the correlation of political polarization on the rhetoric of lawmakers. In political science literature, political word choice analysis has two main roles. First, they feed analyses about larger themes. The words we choose represent our perspectives and inform our audience of our intentions. The way words are chosen and the way issues are framed reflect key aspects of a person's ideals. Second, statistical word analysis allows researchers to evaluate the political context and consequences of rhetorical choices. (Monroe, Colaresi, and Quinn 2008, 372) In a statistical analysis of word choice, Monroe et al. used a corpus of congressional speeches provided by the Dynamics of Rhetoric and Political Representation Project to map and categorize the common words used during floor debates. This study is instructive in the field of text-as-data as it provides a concrete statistical analysis of speeches on the floor of Congress. Monroe et al identify two distinct categories in the field of text-as-data: *feature selection* and *feature evaluation.* Feature selection concerns the selection of words or the ways that each party uses words. This goal is characterized by a binary selection, something is either in or out. This is in contrast to feature evaluation, which not only looks at the words themselves, but also the amount. More than just which words are selected by each party, it evaluates how much each word is used. Monroe et al use a quantitative analysis of both approaches in order to determine the linguistic differences between the speeches of Republicans and Democrats. (Monroe, Colaresi, and Quinn 2008, 374)

In the feature selection evaluation, Monroe et al measured the differences between the two parties with respect to certain words. Specifically, they find that during the floor debates about abortion, Republicans used the terms “baby” and “procedure” more often, whereas Democrats more often used “right” and “women.” Their research demonstrates the value in using text as data to evaluate overarching themes. Their research also provides a large amount of mathematical calculations which provide insight in to the possible ways my research could be structured. For instance, rather than looking at whether a specific word is used more over time, it would be more useful to observe the consistency of the vocabulary certain time periods. Additionally, to eliminate any confounding variables like controversial topics saturating the data with specific phrases, each time period should be measured based on the amount of unique words in each dataset.

Research has also been devoted to the context of the words, not just the words themselves. After all, while both parties may use the word ‘global warming’, each party might use the word with different connotations, meaning that while their language might seem the same, they are actually saying two separate things. A Democrat might be more likely to use the term global warming in a positive light, in saying that we should make effort to curb its influence, whereas a Republican might be more likely to use it in a negative context, saying it is a natural process or nonexistent. With that in mind, a statistical analysis of just words alone may leave out key context which could be relevant to the issue at hand. This is exactly what Box-Steffensmeier and Cannon did when they observed one-minute speeches in the context of party loyalty. Using speeches separated out by congress-member, they then used a statistical analysis tool called the Linguistic Inquiry and Word Count, or LIWC, to capture all the words that were used to refer to the other party during floor speeches. (Pennebaker, Booth, and Francis 2007) The LIWC additionally allows the observer to capture positively or negatively associated words. They found that in eras of high party competition, the words used to refer to the opposing party became more negative. While partisan speechmaking existed in all eras of study, during the 103rd Congress (1991-92), partisan attacks increased dramatically. Whereas in previous classes, Republicans made around 1.4 anti-Democrat speeches per representative, in the 103rd Congress, they made around 4.49 partisan speeches per representative. For the Democrats, they jumped from .63 per representative to 1.53 partisan speeches. (Box-Steffensmeier and Canon 2015, 63) The research of Box-Steffensmeier and Cannon provides two considerations relevant to this research. The first is another potential measure of polarization in the form of party competition and derision. The second is the use of the LIWC, which is a useful tool for analyzing and extrapolating data from text without the use of scripts and coding. It allows analysis not only on what amount of each word are said, but also on the polarity of the words themselves. It will count the amount of positive and negative words, as well as the amount of words with an analytic or emotional tone. This provides several different avenues to explore in the content of speeches. While this study is constructed based on pure word counts and does not necessitate consideration of context, the LIWC could be a platform to establish further research combined with the data gathered for this study.

Expanding on the idea of studying word choice, studies have used language sources to predict political ideology. In a piece titled "Language and Ideology in Congress," Diermeier et al use a predictive algorithm to determine a person's political ideology using a writing sample. Using that person's word choices makes them likely to have conservative or liberal leanings. Unlike other papers, this piece lists the specific website used to download all their data. They examined senatorial speeches of the 101-108th Congress downloaded from *Thomas.gov.* Using these data, they were able to gather a list of words with conservative and liberal connotations. Words like handgun, lobbyist, and disabilities all point towards someone having liberal leanings whereas words like ranchers, embryonic, PAC, and unfunded point to conservative ideology. (Diermeier et al. 2012, 43) This research provides not only a resource to gather data, but also detailed methods which could be replicated during further research.

Demographic Influences

Some text-as-data research is related to the gender of representatives, which could potentially be a confounding variable in my research. Bei Yu of Syracuse University investigated the way gender is related to Representatives' expression of messages and party communication. Her research found that female politicians tend to use more words related to emotions and contained fewer articles than their male counterparts. (Yu 2014, 6) Her methods are particularly relevant to my research, as her question is similar but contains different variables. She uses a corpus of congressional speeches from 1989-2008. With that collection of over 150 million words, she used statistical analysis to determine the most frequent words. The speaker's language style was then calculated using the percentage of words that match a certain language style feature. Yu additionally narrowed the scope of her research by first determining what words are "female" and which are "male." Using studies from fiction writing, she determines which words to look for to indicate differences in speaking style. Applying this idea to the question at hand, it is useful to add a set of buzz words used by each party as an additional metric of word frequency.

Studies have also looked at the influence of race in congressional rhetoric. In Dietrich et al., a statistical analysis of congressional speeches broken down by race found that African-American members of Congress were far more likely to use words like “segregation” and “civil rights” than white members of Congress. African-American members also speak on civil rights issues much more frequently and in much more positive light than their colleagues. The same is true of other non-white demographics. All are more likely than white members of Congress to speak on subjects relating to racial prejudice and civil rights. (Dietrich et al. 2017, 32–33) The presence of racial differences presents another dimension to consider in word frequency analysis. Whether a party has differing amounts of a particular racial group may influence the kinds of words that appear in the analysis.

Chapter 2: Methods

Hypothesis

Alternative Hypothesis: Rises in the amount of ideological difference (polarization) between Democrats and Republicans will correlate with increasing similarity between the word choice of party members.

Null Hypothesis: The consistency of word choice in one-minute floor speeches & five-minute speeches demonstrates no discernable relationship to increasing levels of polarization.

To test the consistency of Congressional rhetoric within parties, I will use a quantitative analysis of words spoken while on the floor of the House of Representatives. Specifically, I will use one-minute speeches delivered on the floor of the House from 1995-2018. Congressional one-minute speeches are a succinct expression of a position or belief; they are used by representatives to communicate with their constituents and, increasingly, to attack the other party. Studying floor speeches reflects the ways that lawmakers and parties are attempting to frame the national dialogue of a particular issue. (Maltzman and Sigelman 1996, 820) The goal of the study is to extrapolate what relationship, if any, polarization has with the distribution of the most common words in Congressional speeches.

This study will use a comparative analysis of the difference between median roll call voting scores of each party and the distribution of the most common words. If the alternative hypothesis holds, it will be demonstrated by strong correlation between the higher differences between parties and the less distribution of common words. If the null holds, this will be indicated by changes in the distribution of the most common words that do not match with changes in the distance between both parties.

Cases and Sampling

The bulk of this analysis will be conducted on speech data provided by the *Congressional Record* which starts in 1995 and goes up to 2018. Every one-minute speech said on the floor of the House between those dates are included in the data for this study. For sampling, each year will be split into two 6-month chunks, with every six-month period consisting of one data point. The data will be split into six-month sections in order to evaluate change from year to year. Dividing the data this way will allow show more detail about the changes in word distribution. Additionally, using data sets of this size will show any changes in word frequency that occur in concurrence with elections or other significant historical events. The data selection will start in January of 1995 and continue until December of 2018. Each 6-month chunk has a word frequency score assigned to that measures the distribution of the most common words. Additionally, another analysis will be conducted based on specific words that previous research identifies as indicators of partisan speech. The frequency score can then be plotted on a line chart with the year as the x axis and the similarity score on the y axis. This chart will demonstrate how the distribution of common words in Congressional speech changes over time. If the line shows a rising trend overall, then word choice similarity is increasing with each data point, which would be in line with polarization. If the line shows a decreasing trend, then that demonstrates that word choice is diverging within a party, which would be contrary to trends in polarization.

Before the data from the *Congressional Record* can be analyzed, the speeches need to be divided into Democratic and Republican groups. The metrics of polarization, this study require a comparison between the parties in order to establish ideological difference. However, the word distribution portion of this study does not require both parties to be considered within the same analysis. In order to gain the most detailed picture of changing word distributions it is best to conduct separate analysis on both parties. Republicans and Democrats will inevitably have subtle differences in the way they express their messaging. If the parties are grouped together, that runs the risk of muting smaller trends and observations that are specific to either party. Separating the speeches into two databases prevents the speech analysis from conglomerating into white noise that provides no discernable results.

Data Sources and Data Collection

The first measurement for this analysis is polarization. The first method of quantifying and plotting levels of polarization is with roll call voting or DW-NOMINATE scores. The data for roll call voting was acquired from *voteview.com*, which provides the raw data that can be analyzed in R to produce the metric that shows polarization trends. The second method of measuring polarization is the Political Polarization Index. The PPI comes from a paper written by Maria Azzimonti, who places polarization on a timeline that starts in the mid-1950s.

For congressional speeches, the speech data was gathered using the *Congressional Record* downloaded from *Congress.gov.* Isolating the individual speeches from the larger corpus of words said on the floor of the House presented a unique challenge for this study. Much of the *Congressional Record* taken strait from

Congress.gov contains many different memos and procedural notes which weren't actually said on the house floor. Additionally, the amount of speeches delivered in the past two decades is immense. There have been thousands of congressional representatives in that time frame, all of whom make dozens of speeches every year.

Fortunately, there are computer programs that were written for studies like this. To isolate the speeches and parse them into readable data I used a python script which automatically scrapes the *Congressional Record* and parses out the speeches into Comma Separated Value (CSV) files which can be easily read by statistical analysis software. (Nicholas, Carbaugh, and Young 2017) The script allows the user to parse out all speeches delivered on the House floor between two selected dates, provided the data exists on the *Congressional Record*. The software functions using the command prompt, Windows PowerShell, or similar programs. The first step is to set the command directory to the file folder where the script is contained and install all the dependencies. The next step is to activate the scripts help information using the command *python -m congressionalrecord.cli -h*. This is optional but will provide useful information on the positional arguments that allow the script to function. The script is operated by typing *python -m congressionalrecord.cli* followed by the dates between which the script will parse. The dates are entered in a YYYY-MM-DD format and do not have any other code proceeding them. The third component is the do_mode argument, formatted as "pg – csvpath" followed by a file path that will serve as the output for the script. When the script is running, folders named for each day will slowly begin to appear in the output directory. The script takes roughly an hour to scrape 6 months of speech data. Each folder will have pdf documents of the *Congressional Record* from the day matching the title of the folder. The speeches will be automatically extracted from the extensions of remarks document and placed into the master CSV file named "speeches." The script also parses out contents of each bill proposed in the House; bills are not relevant to this study but could potentially be grounds for further research.

Midway through my study, I ran into a setback regarding this script. The script drew from the Government Publishing Office FDsys database that provides the *Congressional Record* with all of the content parsed by the script. Just as I began to collect data for this study, this database was retired and replaced with the *govinfo* database. This was a pressing issue, as the script no longer could gather data and needed to be modified to work with the new database. I encountered an issue where the script would stop functioning very frequently, giving an error that explained that it found no data for that particular day. I would have to modify the command to exclude that day and then keep moving forward. Every fourth or five day that I attempted to extract would elicit this error, which made the process of gathering almost 25 years of speech data prohibitively work intensive.

I noticed that the days that produced the error were almost all on weekends or holidays, which were days that Congress was not in session. The script is supposed to skip days that Congress is not in session and move on to the next day, so it was seemed likely that this part of the script was not functioning properly. This may have been a result of the differences between the new and old database. The old database might have an empty file for days where Congress was not in session, whereas the new one has no file at all. This difference would cause an error in the script, as it would attempt to find data and return a null value. The inability to consistently gather data presented a large problem for me as I did not possess enough experience with python in order to fix the error. I consulted with the Nicholas Judd, the co-creator of the script and explained the issue. He informed me that I was likely correct in identifying the problem and gave some advice about possible methods to fix the issue. He also informed me that he would fix it eventually but was not available immediately. I was not sure whether this meant it would be fixed it in week or a few months. With only had three weeks to gather my data, a more immediate solution was needed to guarantee that I could get usable data. I am fortunate to be friends with very talented coders who said they would be willing to assist. Together we were able to create a solution that repaired the skipping function, allowing the script to run normally. This solution allowed me to gather my data for the next few weeks until the Nicholas Judd posted an official patch that fixed the issue.

Measures

For the purposes of this study, there are two key variables: political polarization and congressional word choice. Word choice, or more specifically common word distribution, is the key variable in this study and requires special consideration. When analyzing a large corpus of text with multiple authors creating works of varying length, there are several potential confounding variables that need to be accounted for. First, speechmakers may have varying speaking styles. Some Representatives speak faster than others, which means that they can fit more words in a one-minute speech. This is accounted for by the volume of speech data included in this analysis. Since computer automation has allowed the inclusion of every single Congressional speech, any small variations in speaking style will be negligible when compared to the total size of the body of text.

A second confounding variable is that representatives generally do not use floor time to the same extent. Some representatives may speak more often than others. Some representatives may only make a few speeches throughout their whole term, whereas some are in office for decades and make hundreds of speeches. This potential complication is also addressed by the size of the data set being used. Including every speech over the span of two decades reduces the influence of any one representative's speeches.

A third potential confounding variable is the time constraints of speeches. Not all speeches in Congress are limited in time. For instance, speeches in the Senate are not beholden to any limitations on time or subject matter. Including Senatorial speeches in the analysis of word distribution could cause individual legislators to become over-represented in the analysis, skewing the data towards that legislator's speech quirks. A filibuster on the Senate floor may last for hours and contain thousands of words. This would dwarf any speeches made in the House and would minimize the influence of House speeches in the analysis. This variable is controlled by limiting the sampling of speeches to just one-minute speeches delivered on the floor of the House of Representatives. The House is much more tightly controlled than the Senate, due mostly to its larger numbers. Speeches are limited in time and often in content. That is why one-minute speeches are the best source of data for this analysis. Concentrating on one-minute speeches increases the reliability of studying spoken words since it creates a more uniform sample population. This limitation will ensure that the data base is made up of a consistent sampling groups of around a few hundred words per speech. Using one-minute speeches also controls for the possibility that certain lawmakers may have speaking habits which result in certain words being used disproportionately.

Another factor to consider when examine speeches is the intended audience or intended message of the speech. Speeches on the floor of the House are not just made to persuade the members of the opposing party. Since the invention of television and C-SPAN, floor time has taken on a second purpose where representatives use speechmaking to communicate with voters. For the purposes of this study, one-minute speeches can be put into two loose categories. First, there are constituent communications, where a representative will be speaking for the purposes of communicating with his or her constituents. Second, there are partisan communications, where a representative is attempting to send a message or attack the positions of his or her colleagues across the aisle. In certain situations, there is overlap between the two categories. For instance, a representative may attempt to curry favor with constituents by attacking positions of the other party that will pose an imminent risk to their interests. However, the distinction that is most relevant to this study is the federal versus state divide. Constituent communications are primarily concerned with issues of relevance to the representative's district. This may include such things as honoring the career of a local public figure or calling attention to the achievements of ambitious middle schoolers. Constituent communications are not substantially related to the dynamics between the parties. (Maltzman and Sigelman 1996, 820)

Despite not directly addressing party relations, constituent communications are still included in the study. While these speeches are not the most potent expressions of polarization, the effects of polarization may also express themselves in the kind of constituent communications a representative chooses. For instance, a representative may try to use a story about a noble police officer to make a larger point about gun control or shine light on an immigrant who created a small business to make a point about immigration reform. Even though these communications are intended to be of local relevance, they cannot be completely divorced from the larger congressional landscape. Representatives may use constituent communication to express larger themes relating to national issues. Additionally, this study focuses on words, not subject matter. If there is a relationship between word choice and polarization, then this will affect all speeches made on the floor, not just those made with the intent to be partisan. To exclude constituent communication from this analysis would be to assume that an individual is capable of switching off party dynamic influences. The works of Hunter et al and Fiorina et al suggest that the polarization of Congress arises from subtle social coercion within the party as much as direct voting stimuli. As a result, the psychological influences of polarization should pervade speeches not directly aimed to address party politics.

Partisan communications are the most direct expressions of polarization. This is similar to the distinction that Morris et al used to study the frequency of words in congressional speeches during certain policy debates. They only looked at the speeches which addressed the opposite party in a negative manner or their own party in a positive manner. (Morris 2001, 107) Partisan communications are an expression of the individual's beliefs as well as the positions of the party. These communications will be the most entrenched in federal politics and the strongest expressions of political polarization. Both forms of congressional communication are included in the data for this study. This study includes a large variety of speeches made up of every single representative who spoke on the House floor between 1995 and 2018. Any singular dissonant speech will be insignificant in such a large dataset.

Having established what kinds of speeches will be analyzed, the next step is to establish ways to quantify word frequency. There are two metrics for word frequency that will be used for this analysis. The first metric is taking the top one hundred most commonly used words and putting it as a ratio of the whole body of text. This will provide a measure of the skewness of word distribution and show how concentrated the most common words are. The more that the data is skewed towards the most common words, the more consistent the speeches are. The second axis of comparison is a narrower analysis using a small set of politically charged "meat words" as a ratio of the whole body of text. These words are chosen using suggestions from literature that point to specific words being more prevalent This metric shows what percentage particularly polarizing words like make of all the words.

Each of these approaches has problems that need to be overcome. For the first metric, it is a phenomenon known as "Zipf's Law." George Zipf was an American linguist who observed recurring pattern in the distribution of word usage. Zipf's law states that in any body of text, the frequency of a word will be inversely proportional to its rank on the list of most common words. The result is a chart with extremely low skewness, where almost all the most common words outnumber the rest by several orders of magnitude. (Robinson and Silge 2018) This is demonstrated in *Figure* 5, where the most common words of several Jane Austen novels result in similar charts.

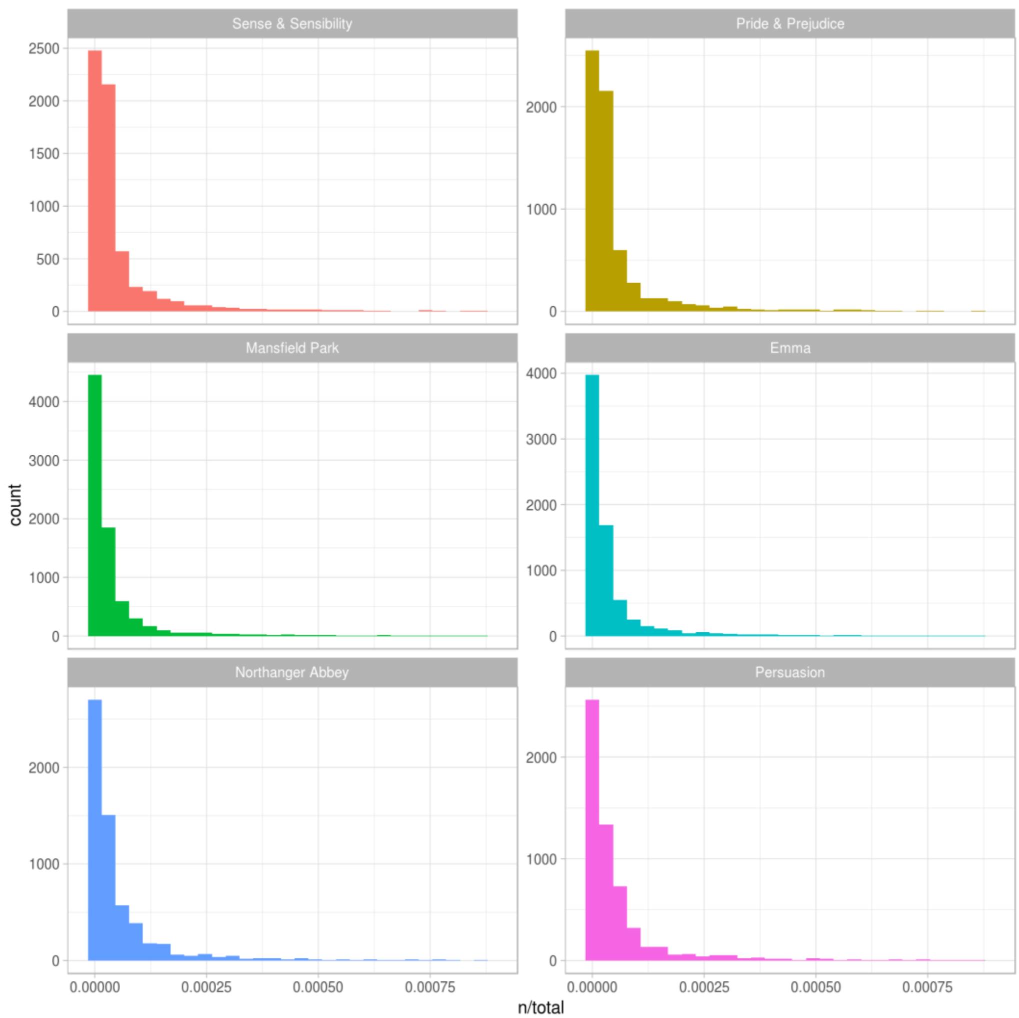

Figure 5: Jane Austen word distribution (Robinson & Silge 2018)

Zipf's Law raises an issue with the initial construction of this project. The first metric of polarization looks directly at the most common words in a set of speeches.

Zipf's Law makes doing analysis on an unedited database unproductive, as doing so would result in each data point having almost exactly the same frequency score. This would create a flat trend line where it is impossible to discern any changes in word distribution. In order to combat Zipf's law, this analysis excluded words that become too common and are normal parts of speech. Eliminating words like 'the', 'when', and other common words allows analysis to be conducted just on nouns, verbs, adjectives, and adverbs. Taking these significant words and weighing them as the percentage of the whole allows for a more accurate way to measure the skewness of word distribution.

The second metric of word frequency uses specific politically charged words that are more associated with the speechmaking of one party over the other. The most immediate problem with this metric lies with deciding which words represent partisan speech. An obvious candidate to create a list of partisan words could be the suggestions of the party message committees discussed in chapter two. However, these are not readily available, so the words have to be found from other sources. One such source comes from a memo sent out by Newt Gingrich in 1995 to members of the Republican Party. The purpose of the memo was to provide Republican candidates running for re-election a list of words to help communicate their message. The memo contains two separate lists: one of positive words, and one of contrasting negative words. Positive words include words like "mobilize," "children," and "pristine" whereas the negative words include "abuse," "decay," and "shallow." (Gingrich 1996) Since this memo was only directed at Republicans, the Democrats required a different source of words. The source for these words came from *Messaging Matters*, a research group focused on how politicians use rhetoric. They provide a list of popular Democratic phrases that include positive words like “bridge,” “future,” and “choice” or negative words like “scam,” “sabotage,” and “drill.” Words from these two sources, Morris et al, and the database constructed for this study all conglomerate to create a short list of around hundred meat words for each party which can be seen in *Figure 6*.

Republican Party		Democratic Party	
Positive	Negative	Positive	Negative
Business	Betray	Activist	Abuse
Challenge	Bureaucracy	Affirmative	Bigot
Children	Bureaucrat	Affordable	Billionaire
Compete	Cheat	Balance	Bizarre
Confident	Collapse	Bridge	Coercion
Conflict	Consequence	Build	Corrupt
Control	Criminal	Change	Crusade
Courage	Crisis	Choice	Cut
Debate	Death	Choose	Damage
Enhanced	Debt	Community	Deficit
Fair	Democrat	Dream	Endanger
Family	Destroy	Empower	Excessive
Fiscal	Disgrace	Energy	Greed
Flag	Excuse	Equality	Hate
Freedom	Failure	Future	Hollow
Help	Hypocrisy	Hope	Immoral
Incentive	Ideological	Humane	Insensitive
Initiative	Illegal	Learn	Intolerant
Legacy	Impose	Listen	Irregular
Liberty	Incompetent	Live	Irresponsible
Life	Insecure	Medicare	Liability
Light	Liberal	Mend	Obsolete
Mobilize	Lie	Movement	Oppose
Moral	Machine	Neutrality	Partisan
Opportunity	Mandate	People	Problem
Passionate	Pathetic	Prosper	Racism
Patriot	Patronage	Prosperity	Republican
Pioneer	Punish	Protest	Selfish
Power	Quo	Provide	Sexism
Precious	Radical	Rebuild	Steal
Preserve	Shame	Reform	Stifle
Principle	Sick	Safe	Strike
Pristine	Socialist	Security	Suppress
Protect	Spend	Share	Torture
Right	Spending	Social	Unconstitutional
Strength	Tape	Society	Unethical
Tough	Tax	Speech	Unfunded
Truth	Taxes	Support	War
Value	Threaten	Undocumented	Waste
Work	Traitor	Welfare	

Figure 6: Republican and Democratic Meat Words

As seen in *Figure 6*, there is a roughly even proportion of positive words and negative words for each party. The words used in the analysis also included plurals for nouns as well as past and present tenses for verbs. For example, if the word is “spend,” then the database is filtered for “spent” and “spends” as well. Each set of words is used to filter the speech data using just the strongly partisan words. This gives a way to avoid the effects of Zipf’s Law. Even if the proportions of the word distribution are the same, the relative position and rank of words may still change. If this study’s hypothesis is correct, then in a more polarized Congress each party’s meat words will be ranked higher on the word frequency distribution than in eras of low polarization.

One notable shortcoming of this analysis is that certain words only become polarized with the presence of other words to put them into context. For instance, the phrase “red tape” is a common Republican phrase, but in this analysis, it would only give the two words separately. However, there are very few other situations where a representative would use the word “tape” during the course of Congressional speechmaking so including just that part will capture the breadth of the issue. The same is true of the word “quo” which is intended to capture the phrase “status quo.”

With word distribution quantified, the next step was to establish trend lines in congressional polarization. Political polarization is much trickier to measure since doing so involves trying to represent ideology with numbers. For this study, there are two main methods for operationalizing polarization. The first is the aggregate DW-NOMINATE scores of each Congressional class. (Poole and Rosenthal 2018) With DW-NOMINATE, finding a way to represent polarization numerically over the years is not straightforward. The most substantial problem for this element of the study was how to assign each year a “polarization score” for each Congressional class. For this analysis, polarization is measured by calculating the difference between the two parties median DW-NOMINATE score broken down by Congressional class. DW-NOMINATE scores are measured on a scale from 1 to -1, with 1 being the most conservative and -1 being the most liberal members of each party. Imagining the two parties as separated clusters of members' voting records graphed with regard to economic/redistributive and social/racial policy positions, the approximate difference between them can be quantified using the rough center of each cluster represented in this study by the median score. To ensure the most reliability, median is used instead of the mean for this metric. Analyses which use means will be more strongly influenced by outliers. Using the mean DW-NOMINATE scores could run into issues when Congressional classes with more ideologically extreme representatives skew the average score of their entire party. Using the median score gives the best approximation of a party's ideological center. Put into R and plotted over time with the ggplot package, the result is an increasing trend line demonstrated by *Figure 7*.

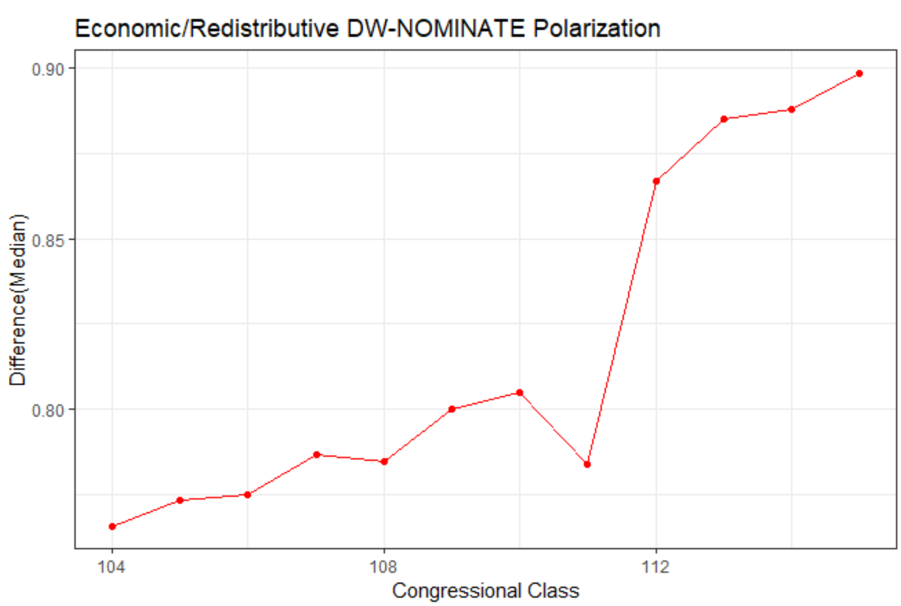

Figure 7: Economic Polarization (voteview.com)

Each data point represents the difference between the Republicans' and Democrats' median DW-NOMINATE scores for each Congressional class between 1995 and 2018. The chart demonstrates that the Economic/Redistributive dimension of DW-NOMINATE has a strong upward trend. In particular, there is a strong upsurge in the 112th Congress from 2011-2013. This upsurge coincides with the 2010 mid-term election, which was characterized by strong Republican gains and Tea Party victories across the nation. In that election, the Democrats lost 63 House seats and their majority, and Congress would remain a divided house for the rest of the Obama presidency. (Graham 2016, 309) Charting the Social/Racial dimension of DW-NOMINATE reveals a similar trend, as shown in *Figure 8*.

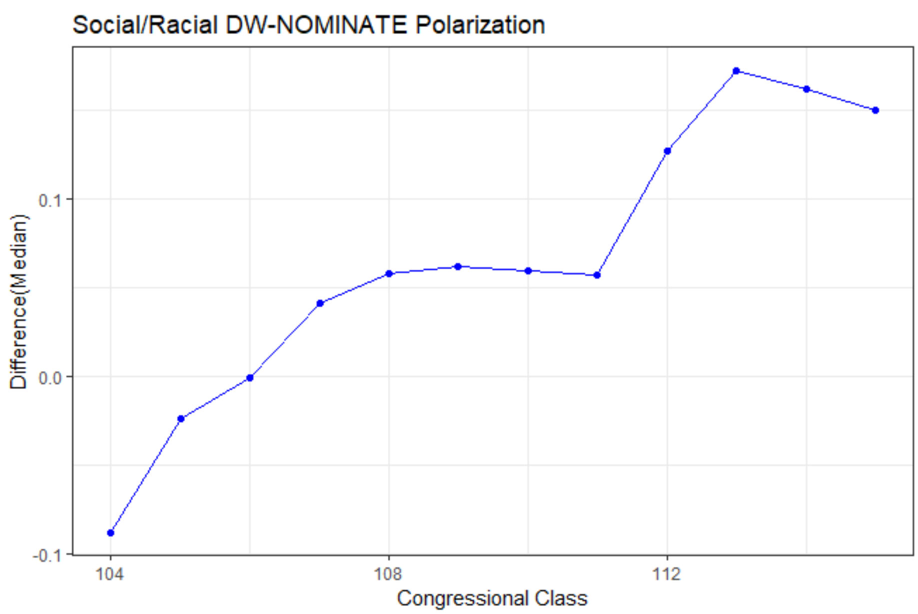

Figure 8: Social Polarization (voteview.com)

Noting the y axis values, the difference between the median values for economic issues is far greater than for social issues. The difference for economic issues ranges from a 0.7 difference to a 0.9 difference, whereas the social aspect only ranges from a 0.1 difference to .15 difference. This observation is supported by scholarship: economics issues have been the prevailing and most substantially polarizing issue in Congress since at least the 1980s. (Wood and Jordan 2017)

The second measure of polarization is the Political Polarization Index (PPI) that uses media responses that mention gridlock and divided government to create a polarization score for each year. (Azzimonti 2013, 5) The PPI is already structured in such a way that lends itself to my question because it is expressed in a chronological plot that progresses one year at a time. The data of this analysis is presented in a similar manner, with each 6-month period of a year being a single data point. While DW-NOMINATE scores evaluate polarization based on the specific voting and legislative habits of each legislator, the PPI gauges polarization based on how outside sources perceive Congress.

Both methods explore the concept from different starting points. DW-NOMINATE scores study polarization by directly measuring the votes of Congresspersons, while the PPI uses public perception of Congress. Combining both methods adds more accuracy to trend comparison. DW-NOMINATE is an inside-out method of measuring polarization that represents the behavior of representatives directly. The PPI is an outside-in metric that gauges the media perception of whether Congress is polarized. Using the confluence of both methods creates a stronger indication of which eras of Congress are truly polarized, which is essential to the accuracy of this study.

Analytic Techniques

The most substantial analysis of this study occurred in determining the frequency of common words in each party's floor speeches. Using the Congressional record scraper, the raw data from the congressional daily digest can be parsed and divided by speech. This was then separated into one-minute speeches, which can be found in the "Extensions of Remarks" section. Then, the Comma Separated Value (CSV) files from that section of the Record were analyzed in R Studio. I chose R as the statistical analysis program for this analysis both for its ability to automate and its flexibility when with reorganizing text databases. The data in the raw database is loosely organized in one long column with the content of each speech making up one row. Each speech also contains miscellaneous designations, code, and punctuation that occur as a result of the scraping process. This makes filtering the database challenging, since there is no separation or breakdown by any relevant distinctions such as party.

For this study it is essential to separate the speeches by party. The parties are used as separate case studies to chart and plot the trends of word distribution over time. Placing both parties into the same dataset runs the risk of any relationship between polarization and word choice being canceled out by the disparate trends of each party. Sequestering speeches into party groups presented a substantial challenge to this study. In the "Extensions of Remarks" section, the *Congressional Record* does not state party affiliation at any point in a speech. This makes it impossible to directly separate each representative into his or her party. However, every speech is consistent in that each begins with the Speaker of the House reading aloud the last name of the speaker. Using this observation, I was able to separate party members by cross-referencing the speech database with a list of party members' names. Sorting the speeches this way required a list of every member of Congress since 1995 -- around 2500 names including repeats. This list was provided by official Congressional resources (Congressional Biographical Directory n.d.) which can be filtered via party and congressional class. The data from the Biographical Directory was transferred into a database with the help of Microsoft Excel macros. I then fed this database into an R script I wrote for this analysis that filtered each set of speech data. The script scanned the speech data for rows that contain strings that matched the names of each representative. The numbers of these rows are marked by the script and then placed into a subset with just the speeches beginning with those names. This will inevitably lead to some overlap, as there are more than a few 'Smiths' on either side of the aisle. However, since their speech data will be included in both case studies, any influence they will have will be balanced by their influence on the other party.

After the parties are separated, I moved on to statistically analyzing each to establish word distribution. Starting with the top 100 words metric, the first step is to use the Text Mapping (TM) R package to tidy up the speech data, this includes removing punctuation, stop-words such as 'the' or 'when,' numbers, and words that will by necessity be in each speech. For instance, every speech begins with the phrase "Mr./Mrs. Speaker," so "speaker" is one of the words that needs to be scrubbed from the database. The data from the script will often contain long file paths or code designations with interspliced letters numbers and punctuation, when all the numbers and punctuation are removed, what results is a long cluster of letters like "crecptpghmr." Fortunately, these strings tend to repeat themselves from speech to speech and are thus easily removed through the same text mapping technique.

After tidying up the database, the next step was to set the corpus of speeches as a document term matrix and create a frequency database using the column sums sorted highest to lowest. This results in a dataset of speeches where the most common word is at the top and the least common word is at the bottom. From this point, the top 100 words can be easily placed into a subset and calculated as a ratio of the whole frequency dataset. For the meat words metric, the process is similar, only that in the word frequency analysis, the list of meat words is put in a subset by identifying the row names that match each word. From this point the process is the same: the sum of the meat words is placed as a ratio of the sum of the whole.

Perhaps the most pressing question is how to display word choice into a "frequency score." For this study, word choice frequency scores are determined using the ratio of the words compared to the body of words as a whole. This approximates the skewness of the word distribution and how concentrated the most common words are to the higher end. Larger skewness indicates less word consistency since the distribution of words will be more even. The percentage of the top 100 most common words or meat words relative to the body of text as a whole creates the frequency scores for each data point. Each metric of word frequency will be plotted and tracked individually, although both analyses will be conducted on the same dataset.

Additionally, the word distribution can also be represented visually. The charts found in Monroe et al's study of congressional word choice are instructive on how to display word frequency. They place each word used on the floor onto a scatter plot where the x axis is the frequency of the word and the y axis is the proportion of the speech data a word makes up. (Monroe, Colaresi, and Quinn 2008, 337) For this analysis, the main purpose of demonstrating word frequency is to show its change over time within a party, a purpose for which the charts by Monroe et al cannot be directly applied. However, they use the same approach where the frequency of a word is scored as its makeup of the whole body of text. Instead, the best chart is a line graph where each data point represents one party's word frequency score for a six-month period. Using every congressional class since 1995 sampled in 6-month periods will produce 46 data points to establish a trend line that shows the change in word frequency in sufficient detail to establish an accurate comparison.

Expectations

The pressures that party members feel to acquiesce to party orthodoxy arise from institutional pressures to curry favor with the party leadership and other colleagues. These pressures create a self-fulfilling process where legislators are continually encouraged, whether through direct or indirect means, to express sentiments similar to their colleagues. Them acquiescing to the expectations and norms in turn reinforces that status quo which puts further pressure on other legislators to adopt similar sentiments in an effort to remain in good standing with the party. I expect that in times of greater polarization legislators will have more similar speechmaking patterns and thus more common word frequency. The presence of greater party pressures will increase the influence of the party message platforms. Representatives will be more receptive to the influence of their party's messaging committee and will give speeches more in line with its suggestions. When speaking or using floor time, party members will be more likely to express sentiments similar to their colleagues on the same side of the aisle.

More party unity will result in more similarity in word choice and thus similar rhetoric. To reach this conclusion, each party would need to have significantly larger proportions of common words in times of high polarization than in times of lower polarization. With the top 100 words metric of word frequency, congressional classes that have a larger difference between party medians will correspond to less skewness in the distribution of words. The top 100 most common words will make up an increasing percentage of the whole as Congress becomes more polarized. This indicates that the more representatives are choosing similar words and similar topics of speech within their party. This will show that word choice is more consistent in a more polarized Congress. For the meat words metric of word frequency, the rank of the meat words will shift to the more frequent end of the distribution, indicating that those partisan words are used in greater numbers when the parties are more polarized. It is more likely that the negative meat words will see the most substantial increases in frequency, since the research of Box-Steffensmeier and Cannon indicates that higher polarization results in a more acerbic relationship between party politics and party speech.

The issue of causation is also relevant to this research. The aim of this study, given its construction, is not to prove that polarization causes more similar rhetoric, but that the development of both are correlated. While correlation does not guarantee causation, that does not disqualify causation from being an important area of study. The exploration of causation between word frequency and polarization is a potential area of further research on the topic of polarization. One potential method of establishing that polarization is the cause of more similar speechmaking would be to conduct a study on the influence of an ideologically consistent environment on test subject's word choice. An experiment could be constructed to expose a subject to consistent speech and then observe how that impacts the content of a provided writing sample. With enough external validity, such a study could have strong implications about the mentality of interparty dynamics. Another approach could be a theoretical one. Evaluating the relevant literature on the nature of group-think could provide support to the notion that polarization causes more consistent rhetoric within ideological groups. While more in the realm of psychology than political science, the convergence of these areas of study would create a new understanding about the innerworkings of our highest legislative body.

Chapter 3: Results

Case Study 1: Republican Party

Once the methods outlined in chapter two were implemented, the results yielded thee groups: First, the top one hundred words, which illustrates how much of the whole body of text is concentrated in the most common words. The second is the proportion of positive meat words, which shows how usage of strong partisan words changed over time. The third is the proportion of negative meat words, which also used preselected words to filter the speech data.

The results are placed onto line charts which show the year on the x axis and the frequency score on the y axis. The frequency score represents the percentage that the targeted words make of the whole. These charts are compared to the DW-NOMINATE and PPI charts to establish the convergence between the trends in word frequency and polarization. Finally, party control and presidential elections are marked on the frequency charts in order to evaluate correlation between significant political shifts and word distribution.

Top 100 Words

The most frequent Republican words were almost always the same few words occasionally switching places. Among these words, the most common was most often “people,” followed by “years,” “today,” and “support.” While the most common words did shift from year to year, the very top of the frequency list was almost always one of these words. Among other interesting words that frequently made the top one hundred were “community,” “family,” and “service.” For the most part the top words stayed relatively the same, with words like "budget" and "spending" increasing in frequency in the former half of each year. One of the lists of the top one hundred common Republican words from the second half of 2018 can be found in *Appendix III*.

In the case of the top one hundred most frequent Republican words, as seen in *Figure 9*, the word frequency varies substantially from year to year. From year to year the scores jump up and fall down creating a saw-toothed graph that reveals no consistent trend line. There are two peaks in the chart, one in the late 1990s and the second in the late 2000s to early 2010s. In between both peaks, the word frequency scores are lower and tend to change in an erratic fashion. The inconsistency of the overall trendline may indicate that congressional word choice is a varied and complex concept that is related to more than just polarization.

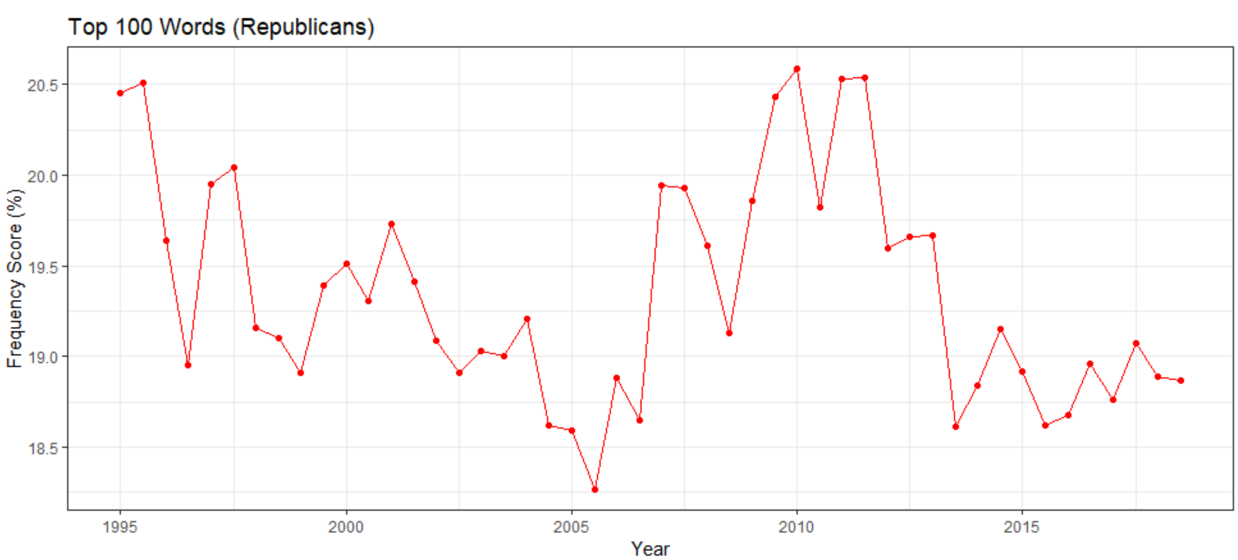

Figure 9: Top 100 Republican Word Frequency

Comparing the trends in the top one hundred Republican words to the DW-NOMINATE metric of polarization, the data on the top one hundred most frequent Republican words does not support the hypothesis that as polarization increases there is a strongly correlated increase in the percentage of common words. In order for this to be supported, there would need to be a consistent upward trend similar to the increase in polarization demonstrated by *Figure 7* and *Figure 8.* Instead, these data show substantial variance that is inconsistent with polarization trends. Beginning at a high point in 1995, the trend line continues downward until it reaches 2005, where the speechmaking of Republicans reaches its least consistent point. At that point, the top one hundred words make up just over 18% of the whole. While the data do not represent a tidy upward slope, there is immense fluctuation in the distribution of the most words. The line begins in a downward slope from 1995 to a low point in 2005. Particularly after 2001, the word consistency among Republicans saw a sharp decline of almost a whole percent over four years.

Next comparing the top one hundred Republican words to the PPI metric of polarization shown in *Figure 1*, the trend does appear to show some correlation. The PPI shows a sharp rise in polarization starting around 2007-2009 that continues to rise until around 2013 when it drops off sharply. This same trend is evident in the top one hundred most frequent Republican words. The trend line turns upward after 2005 and continues until 2012 when it drops sharply. One difference between the top one hundred words and the PPI is that while polarization drops in 2013, it does not drop to levels close to where it was before it rose in the late 2000s. This drop is what happens in the top one-hundred most common Republican words. After the spike in 2005, it returns to levels at or below pre-rise levels. Additionally, according to the PPI, from 1995 to 2010 polarization reaches its peak with the re-election of George W. Bush in 2004. The data for the top 100 words shows the nadir of word frequency at this point, which does not support the notion that polarization is tied to word frequency. All in all, while there are some similarities in the trends of the top one hundred words and polarization as defined by the PPI, those differences are not quite convincing enough to rule out their occurrence by chance.

As the data came in, it became apparent that the fluctuations in word frequency were fairly minute from year to year. Word frequency ranged from a high point of around 20.5% to a low point of around 18%. From year to year, the frequency scores of each party varied by only small portions of a percent. When dealing with such large blocks of data, it is not surprising that the fluctuations will be small. However, this does not mean that the results are not significant. While the percentage changes are small, these small differences equate to large amounts of words. For instance, from late 2006 to 2007, the top one hundred frequency score jumped from 18.27% to 19.94%. Each 6-month speech-data span contains around 2 million words, so a 1.67 percent difference comes out to around thirty-five thousand more identical words in 2007 than in 2006. In perspective, thirty-five thousand words is approximately ninety single-spaced, twelve-point pages of text. While the variations in actual percentage are not wide in their sweep, large changes in word frequency are needed to influence a body of text with over two million words.

While the correlation with polarization is not especially convincing, there are more ways to evaluate the data. One potential avenue is to look at what relationship word distribution has with party control in the House of Representatives and which party holds the presidency in order to observe shifts in political power. This is demonstrated by *Figure 10*, where party control and presidential elections are marked:

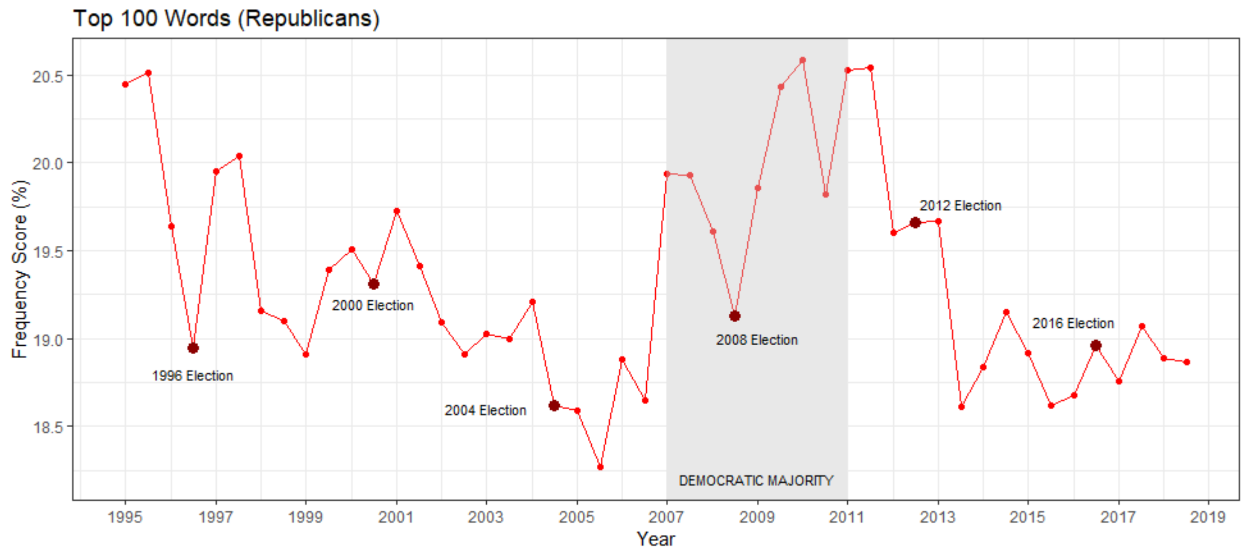

Figure 10: Top 100 (R) with Party Control and Presidential Elections

In this study's timeframe, all but one of the House Congressional classes have had Republican majorities. The only Democratic majority was in the 110th Congress (2007-2009) when the Democrats retained a steady majority. The Democrats gaining a majority in Congress corresponds to a substantial rise in the concentration of common Republican words. From the end of the Republican majority in late 2006 to the end of Democratic control in 2011, the percentage of the top one hundred most frequent words jumps by over a percent. When the Republicans regained control of the House in the 112th Congress (2011-2013), the word distribution returned to the level where it was prior to the Democratic majority.

While polarization is not strongly correlated with the distribution of the top one hundred most common Republican words, the top one hundred common words does suggest that speech consistency has a relationship with party control of the House. A sharp decline in the concentration of common words occurs after 1995 which also deserves attention. This decrease may be explained by the party control of the House. While speech data prior to 1995 is not available from *The Congressional Record*, 1995 was the first year that Republicans gained control of the House of Representatives in forty years. In the years just prior to 1995, the Republican Party was the minority in the House, which corresponds with the high starting point of the chart immediately followed by a sharp drop in common words. This suggests that while the consistency of word choice might not be directly tied to the rise and fall of polarization, it is at least correlated to whether a party is in control of the House. This may be due to the fact that being the minority party increases party unity. (Lenchner 1976, 594) When a party is in the minority, there is more incentive to rally behind the party platform and less incentive to bicker among other party members. Having a party be in the minority could create an environment where representatives put aside interparty differences and focus on regaining control of the House. However, this explanation is not perfect. For one, it assumes that inter-party conflicts will be left behind the moment a party loses their majority. I do not see this as a safe assumption, as inter party conflicts may play a large role in which party retains their majority in the first place, and fundamental conflicts of ideology are not often dropped at a moment's notice.

Another observation that could explain the variance relates to the existence of contentious elections. For instance, the election of Barack Obama in 2008 corresponds with a substantial jump in the top Republican words. Shortly after the 1996 election where Bill Clinton was re-elected, the consistency of Republican speeches also substantially increased before falling sharply. In the case of the election of George W. Bush and Barack Obama, both show similar patterns. When both were elected for their first term, the percentage of the top one hundred Republican words saw a sharp rise. Upon re-election, the opposite was the case, after the reelection of Bush and Obama the Republican word frequency dropped by around 0.5% to 1%. The 2016 election of Donald Trump is a notable exception to this observation. Despite the contentious nature of the political climate of that election, there is no corresponding spike in percentage, rather for the past three years, the word consistency in Republican speeches has remained relatively low. There are many potential explanations for this. For instance, increased disagreement among party members and factions could cause variation in party rhetoric despite having a strong majority.

Positive Words

The next metric of polarization is the frequency of specific partisan words, or meat words. Filtering the speech database using just the positive Republican words and then taking those words as the percentage of the whole produces a trend line demonstrated by *Figure 11:*

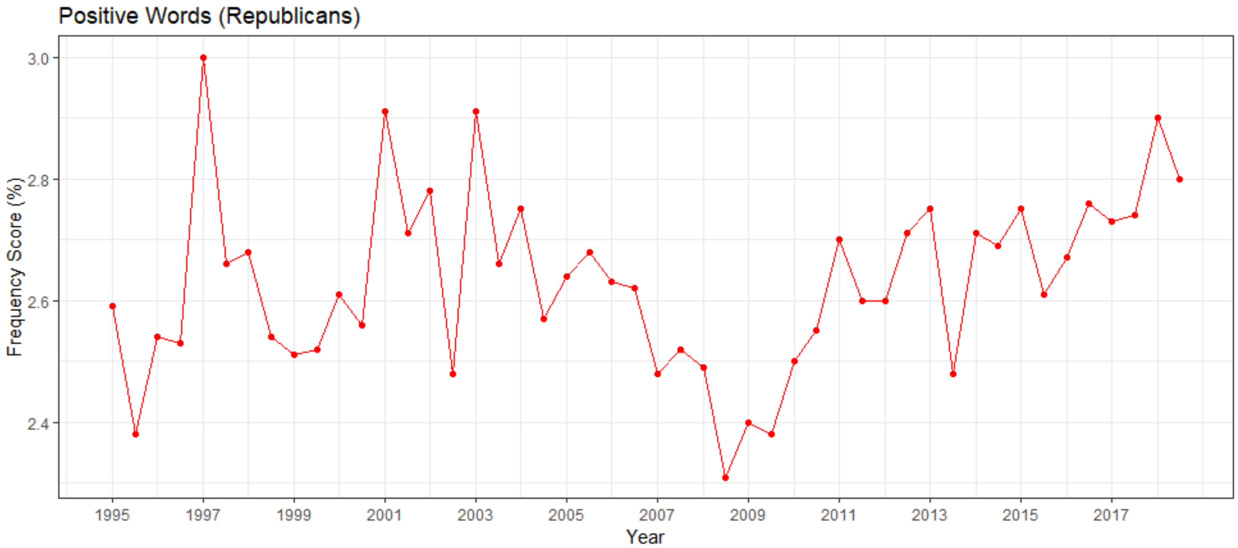

Figure 11: Positive Republican Word Frequency

The positive words metric shows a trend that is almost a mirror image of the top one-hundred words analysis. Rather than reaching its peak in the 2007 to 2009 timeframe, that is when the trend dips substantially. Rather than trailing off after 2011, the trend grows in a narrow wedge to a peak in 2018. This rising trend indicates a correlation with the DW-NOMINATE metric of polarization shown in *Figure 7* and *Figure 8*. Both charts show a steady increase in value starting at the 110th and 111th Congress that continues to rise until the present day. That said, the word frequency from 1995 to 2007 shows much higher levels than the other parts of the graph, which is not supported by the DW-NOMINATE measure of polarization.

Looking at the PPI, the positive Republican words analysis shows a stronger correlation between polarization and word distribution than the top one hundred words analysis. For one, the positive Republican words analysis shows a consistent upward trend starting around the mid-2000s. The PPI and the positive words both have sharp drops in value in 2013 that appear to be discordant with the rest of the data. Furthermore, the positive Republican words trend and the PPI show higher values around the re-election of George W. Bush from 2004 to 2008 that drop before steadily rising again during the Obama presidency. Overall, the positive Republican meat words analysis shows suggests a stronger correlation between Congressional polarization and increasing word frequency than the top one hundred words.

Turning to correlation with House majority and presidential elections, *Figure 12* provides more support to the notion that party control is correlated with changes in the consistency of Republican word choice. When a Democratic majority took office in 2007, there was a substantial drop in the use of the positive words used in this study.

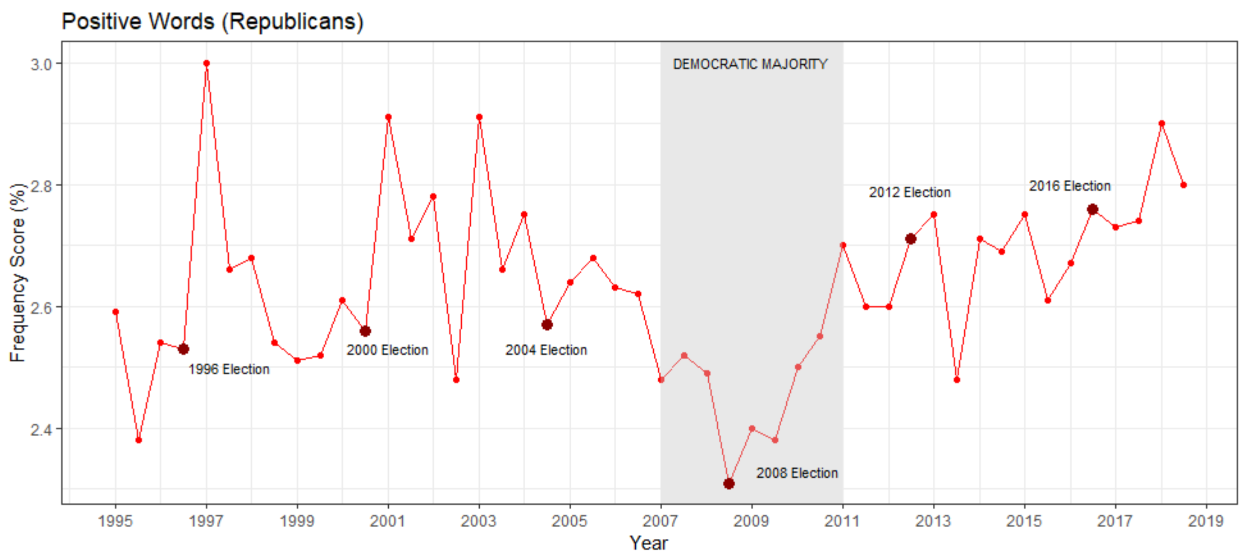

Figure 12: Positive Words (R) with Party Control and Presidential Elections

This result is the exact opposite of the chart of the top one-hundred words, which saw a sharp rise in word frequency during Democratic control. This result seems to indicate that while Democrats are a majority in the House, Republican usage of key positive words decreases. Looking at the 104th Congress in 1995, right after the Republicans gained back control after forty years, the frequency of positive words starts out at an extremely low point, almost identical to the frequency scores in 2008. The top one-hundred words analysis showed the same synchronicity between the two eras in this timeframe where Democrats had a majority; this supports the notion that the word frequency of Republican congressional speeches is correlated with whether the party is in control of the House.

Regarding the effects of presidential elections on word choice, the positive Republican words offer interesting but less conclusive observations. For the re-election of Clinton and the election of George Bush, the word frequency surges upward for the first term after the election before crashing back down to the same pre-election level just in time for the next election. As the elections pass, the upward surge in frequency becomes less and less prominent. After the re-election of Bush, the word frequency experiences a slight uptick. When Obama was elected in 2008, the word frequency among Republicans experiences a small increase that is followed by several years of incrementally higher frequency scores. This trend continues to the 2016 election, where the trend turns in the opposite direction. After the election of Donald Trump, word frequency saw a slight decrease followed by a small jump before tempering back to the same frequency as before the election.

The trend line shows most strongly that positive Republican word choice is related to whether the party is in the majority in the House. This observation has interesting implications for political analysis. These finding give more support to the idea that when Republicans are not the majority party, they are more inclined to go on the offensive and use less floor time to speak favorably of their own party or the state of the nation. As for the presidency, there is no indication that presidential elections of either party have a consistent effect on the positive Republican word choice. While in some years the election of a new President saw a stark jump in frequency score, other years the effect was less pronounced or even the reverse.

Negative Words

The last metric of word frequency is the negative Republican words which include language which may be used to attack the opposite party or policies they disagree with. Putting those words as a proportion of each six-month speech set produces a trend line demonstrated by *Figure 13:*

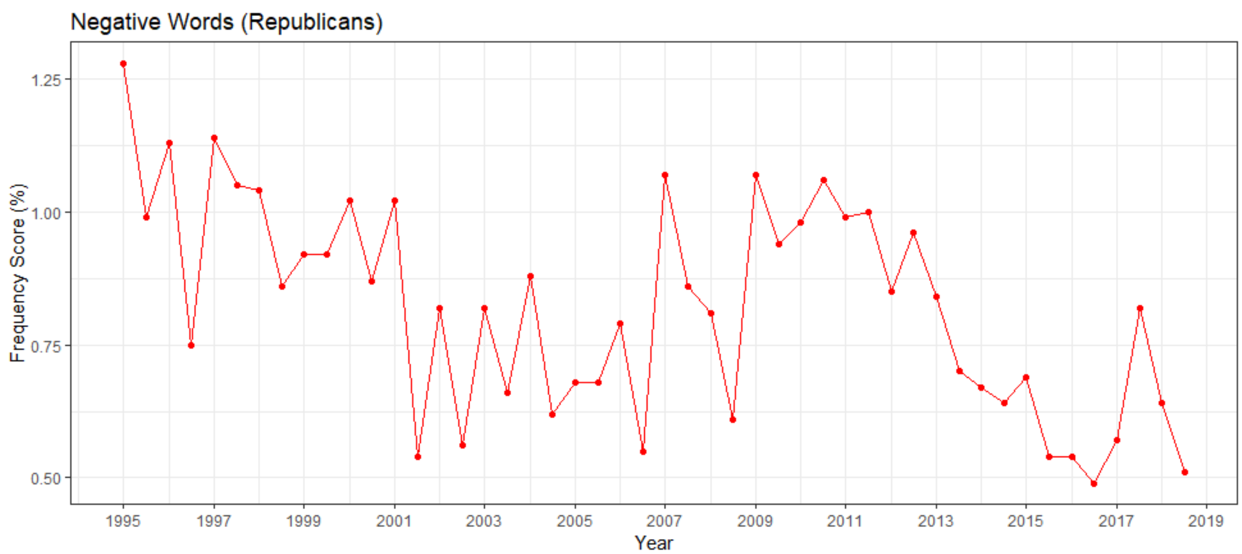

Figure 13: Negative Republican Word Frequency

The trend line begins at its zenith in 1995 and jumps up and down erratically for the majority of the data frame. There are very few consistent trends to extrapolate from these data. However, one apparent feature is two distinct eras of high negative word frequency. The first is 1995 to 2001 and the second is 2006 to 2012. These are the two points in the chart where the frequency of negative words hovers around 1%; in all other points, the score is far below 1%, at times even reaching below 0.5%. After trending upward from 2006 to 2012, the percentage then dips substantially, and barring an anomalous datapoint in 2017, remains very low for the remainder of the data frame. Comparing this trend line to the DW-NOMINATE chart shows very little positive correlation. The DW-NOMINATE chart is characterized by a consistent upward trend that reaches its peak on the most recent data point. In fact, the negative Republican words demonstrate a negative correlation with polarization. The frequency of negative words is almost entirely the opposite trend as DW-NOMINATE. The high point is on the oldest point on the chart and despite several upward spikes still creates an overall downward trend. The lowest point on the graph is also the most recent in 2017-18, which is the opposite of the DW-NOMINATE trend line. Additionally, the only point where the DW-NOMINATE trend takes a dramatic dip is in the 111th Congress from 2009 to 2011. The exact opposite occurred with the negative word analysis, from 2008 to 2011, the frequency of negative words sharply jumped until turning downward again close to 2012. This metric of word frequency provides little support to the notion that DW-NOMINATE polarization is correlated with word frequency.

Turning next the PPI, the negative Republican words are also at odds with the trends of polarization. The PPI shows a strong surge in polarization around 2003 to 2004 that dips below the average before rising again in the late 2000s. The negative words analysis has part of that trend, where the frequency of negative words spikes around 2006 and continues to rise until 2013 when it comes back down again. However, the spike in polarization surrounding the 2004 election is not represented by this trend. The point where the PPI shows rising polarization in the early 2000s matches with a low point in word frequency in *Figure 13.* The point where the negative frequency scores are highest are in 1995, but the PPI identifies this as an era of below average polarization. This metric of word frequency demonstrates little correlation with either metric of polarization, so thus does not strongly support the hypothesis of this study.

Turning to control of the House and Presidency and its possible correlation to Republican word choice, the data of this metric offer more support to the hypothesis that control of the House has influence on Republican speechmaking, as demonstrated in *Figure 14:*

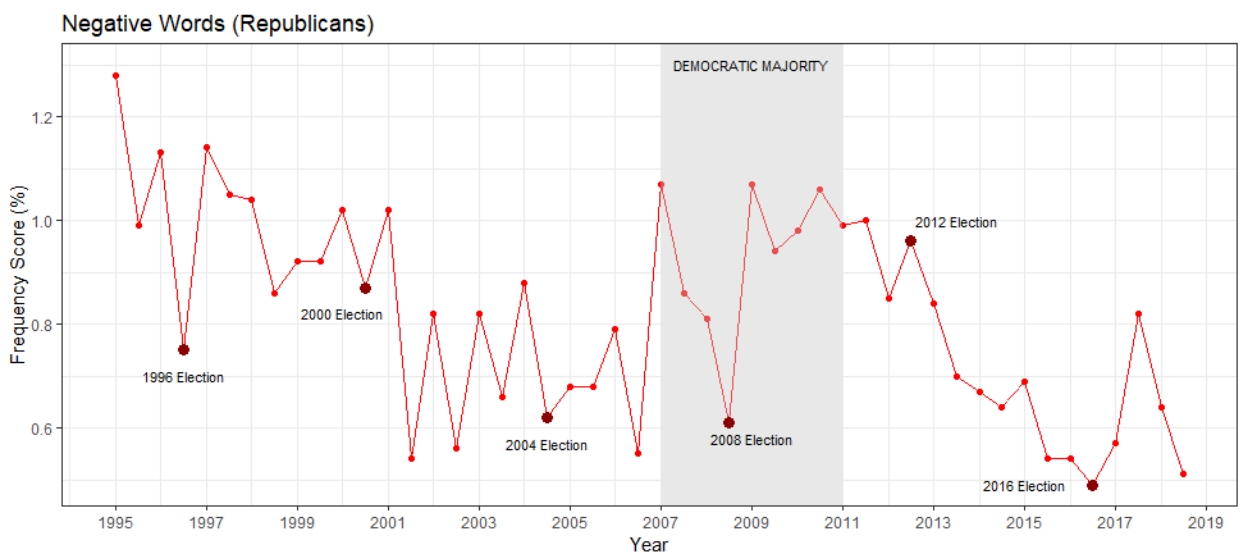

Figure 14: Negative Words (R) with Party Control and Presidential Elections

The negative Republican words metric shows that the Republicans use negative meat words more frequently when they are not in control of the House. In 1995 when the Republicans regained control of the House, the word frequency dropped from a high point of almost 1.3%. The word frequency scores continued to trend downward until

2007 when Democrats regained a majority. The era where Democrats were in control is at the peak of an era of rising negative word usage among Republicans. Barring a two-year dip between 2007 and 2008, the commanding trend surrounding the Democratic takeover of the House is a substantial rise in Republican negative words. With positive words, it was the case that losing a majority in the House was correlated with a decrease in positive word usage. In this chart, the opposite is true: not having a majority is correlated with an increase in negative word usage. Combined with the positive Republican words analysis these data support the observation that control of the House chamber corresponds with a rise in negative language and a decrease in positive language.

Looking next at correlation with presidential elections, the pattern of negative Republican words has a similar pattern to the positive words. Starting with the 1996 election, the frequency of negative words sees a substantial decrease during election season and then an equally impactful increase in the Congress following the election. The same pattern repeats itself for the next election cycles, where the 6-month period before an election sees a decrease in negative word frequency and the first session after the election sees a return to pre-election levels. The pattern becomes the strongest during the 2008 election where the frequency score drops below 0.6% before jumping up to over 1% the following session. The only instances where this pattern does not hold is in the 2012 election, where the trend shows a decrease in word frequency. In the term prior to the 2012 election there was a substantial increase in the usage of negative meat words that was followed by a downward trend that continues up until the 2016 election, where the pattern returns to levels similar to before 2012. The pattern of a dip in word frequency surrounding a presidential election is repeated in both Republican meat word frequency metrics. In both cases, the 2012 and the 2016 elections appear to represent a break in the trend, before returning back to previous levels. This pattern seems to suggest that when there is a presidential election, the word frequency of congressional speeches becomes less similar.

Case Study 2: Democratic Party

Top 100 Words

The top one hundred most common words used by Democrats had similar words to Republicans. Just like with the Republican top words, "people" was the most frequent word during most years. It was joined in the top words by "president," "community," "support," "act," "colleagues," and "work." The top one hundred Democratic words also often included words like "justice," "health," and "program." One six-month data segment of the top one hundred Democratic words from the ladder half of 2018 can be found in *Appendix IV*.

Just like with the Republicans, the top one hundred words among Democrats shows no consistent rising or falling trend throughout the data frame. While the top 100 words among Republicans showed consistency in the form of two distinct eras of rising word frequency, the Democrats do not share this trend. The Democratic trend line starts at a high point in 1995 and then falls down below 18.5%. From there it fluctuates up and down repeatedly and never reaches or even comes near the frequency score in 1995. The trend also does not have distinct eras where it is rising and falling, the majority of the trend fluctuates by large amounts every two to four years, as demonstrated by *Figure 15*.

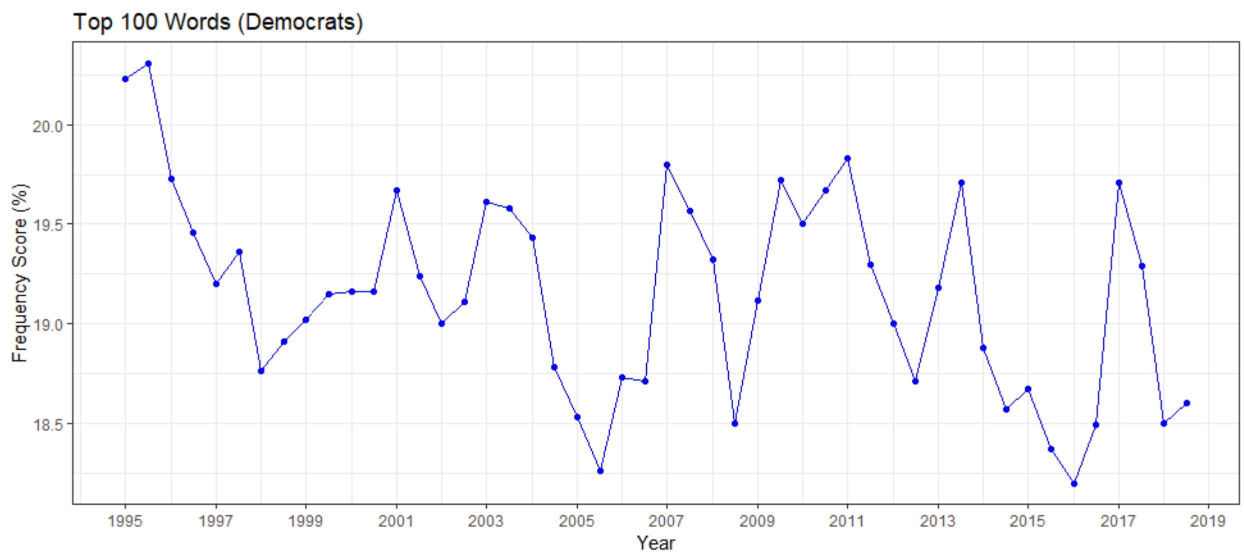

Figure 15: Top 100 Democratic Words

Just like the top one hundred Republican words, these data do not have a steady upward or downward trend. Based on the DW-NOMINATE analysis of polarization, the top one hundred Democratic words does not support the hypothesis that polarization is correlated with word frequency. Unlike the DW-NOMINATE metric, the top one hundred Democratic words stay relatively consistent if the values are averaged out. Though there is substantial fluctuation between election cycles, any rise is counterbalanced by an equivalent drop in word frequency after the election. This provides evidence against the hypothesis that polarization and word frequency in Congressional speeches are correlated; if a correlation were present it would be expected that the trend line of words frequency would be at its high point in 2018, which is not represented by this metric.

When compared to the PPI measure of polarization, the top one hundred Democratic words shows a small amount of correlation between polarization and word frequency. The PPI shows polarization at its two most extreme moments around 2004 and 2011. In the top one hundred words, both of those eras match in terms of increased values. These data show increased word frequency surrounding 2004 and also surrounding 2011. Additionally, both years are flanked by substantial drops in word frequency, just as the PPI shows polarization dropping around 2006 and 2013. The PPI and this metric are at odds, however, when it comes to the pre-2000s era. The top one hundred words shows 1995 through 1999 as an era of high concentrations of words with a strong downward trend. Taken as a whole, the PPI and the top one hundred Democratic words show a weak correlation.

When party control and presidential elections are placed onto the chart area, it shows some interesting relationships between elections and party word frequency. The first observation is that the Democratic majority in the House from 2007 to 2011 seems to have little correlation with the Democratic word distribution during floor speeches, as demonstrated by *Figure 16.*

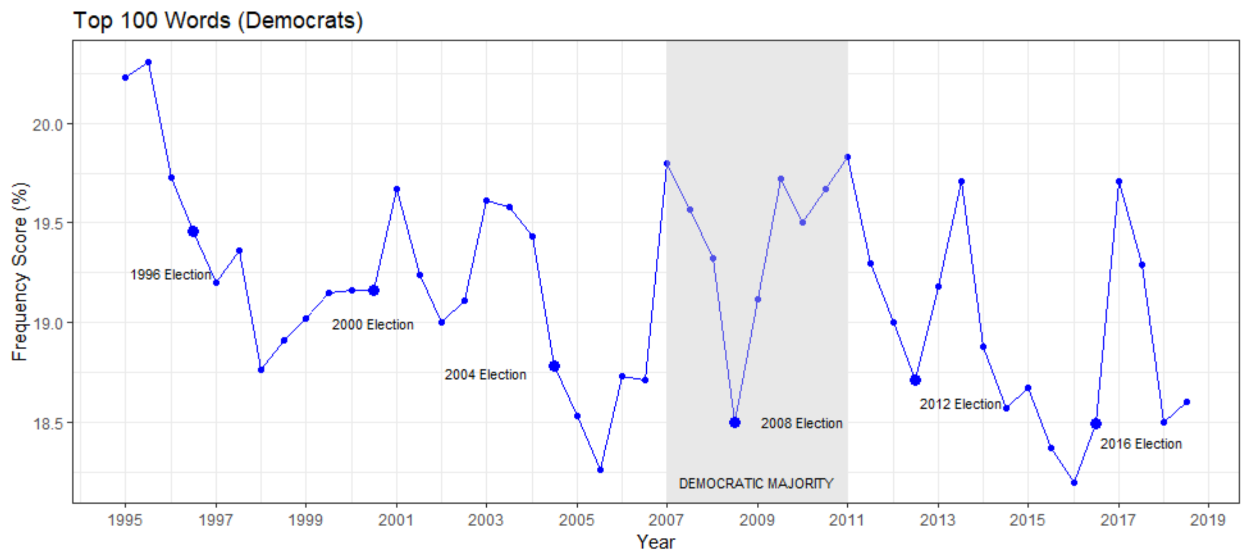

Figure 16: Top 100 (D) with Party Control and Presidential Elections

While both the year the Democrats gained majority power and the year, they lost it have very high word frequency, most of the times in between saw the same word frequency fluctuation as other eras in the data frame. The extreme fluctuation in party messaging even in eras where they are the minority party could potentially suggest that the Democratic Party is far less disciplined in its rhetoric than the Republican Party, or that they address a wider array of topics and policies in speechmaking. Either way, it is fascinating that the Republicans word frequency much more strongly affected by the influence of party power dynamics.

The word frequency of the Democratic Party appears to be much more strongly correlated with Presidential elections than who controls the House. Barring the re-election of Clinton in 1996 and Bush in 2004, each election follows a similar trend: the frequency of common words drops in the year of the election, and the Congressional session immediately following the election is marked by an extreme rise in word frequency. The two most extreme examples of this trend are in the 2008 and 2012 elections of Barack Obama. In 2008, the 6-month period prior to the election saw a dip in frequency from around 19.25% to around 18.5%. The following election saw similar numbers with a percentage drop just before the election and a sharp rise after the election. The strongest rise in word frequency happened after the contentious 2016 election, where the frequency score rose from 18.5% to around 19.75%. These data suggest that the word frequency of the Democratic Party has a stronger correlation to who is control of the presidency that with who is in control of the House. The repeating ‘v’ pattern surrounding elections is also a fascinating trend. One could speculate that this trend is caused by the individual representatives becoming more focused on their own election than on the party message. Around election time there may be more incentive to spend time on constituent communication and district specific speeches.

Positive Words

The positive Democratic words illustrated in *Figure 17* reveal a trend line with eras of fascinating consistency flanked by eras of extreme fluctuation.

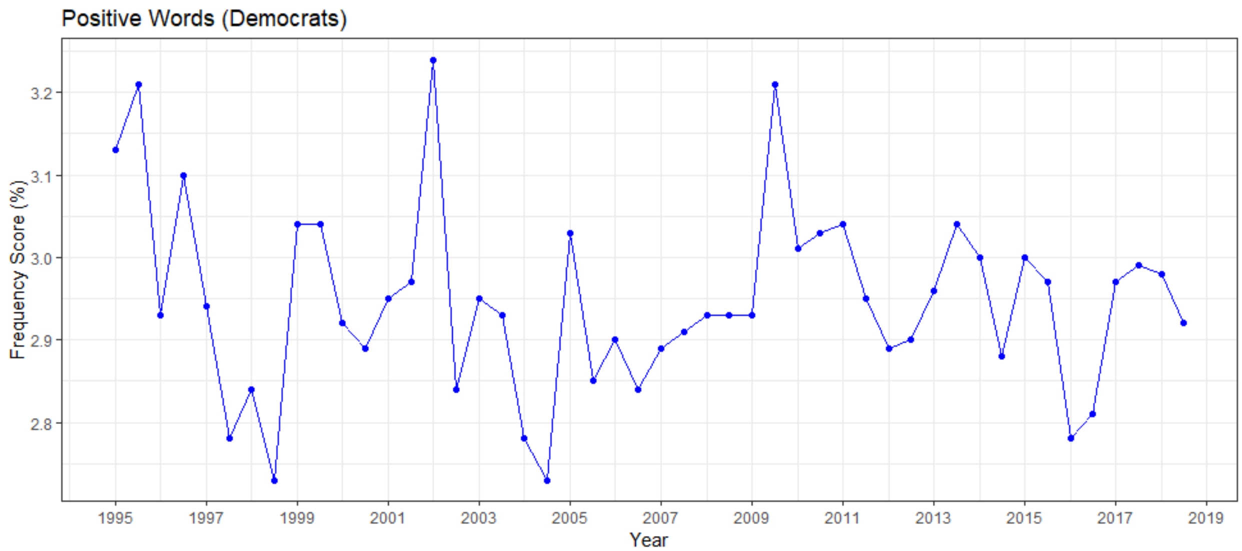

Figure 17: Democratic Positive Word Frequency

Just like the top one hundred words, the highest point is right at the beginning in 1995. From there word frequency immediately drops sharply. This drop continues into the late 1990s where the trend again turns upward until the early 2000s. The mid-2000s sees another substantial downward turn in word frequency that lasts until around 2009, where the third of three dramatic spikes occurred. The trend then continues at an elevated level with a slight downward trend until 2016. The three spikes in the data occur at 2002, 2005, and late 2009 and is followed by an era of high frequency and a downward trend. Between 2007 and 2009, the positive word frequency remained remarkably consistent, with a variation just over 0.05%.

When compared to DW-NOMINATE polarization, the positive Democratic words show little correlation. Just like with the top one hundred words, the trend line is not beholden to any long-term trends. Rather than getting better or worse over time, the trend is inconsistence, with some years having much higher frequency than any of the years around it. Time periods that DW-NOMINATE identifies as eras of high polarization are marked by this metric as eras of low word frequency. The year of lowest polarization in 1995 is the year of the highest word frequency. This metric provides little support to the notion that polarization in correlated with an increased rate of word frequency.

There is a similar lack of correlation between word frequency and the PPI. The PPI's most significant eras of high polarization correspond with eras of low word frequency. While the rising polarization shown from 2007 to 2011 does match with a period of strongly increasing word frequency, the 2003-2005 era of high polarization is not reflected by this metric. When compared to the other metrics of word frequency, the positive Democratic words demonstrates little correlation with polarization when compared to other metrics. As a whole, the Democratic word frequency appears to show less correlation with polarization than does the Republicans.

Turning next to examine possible correlation of word distribution to the control of the House and Presidency, a similar trend is shared between the positive Democratic words and the top one hundred Democratic words, demonstrated by *Figure 18*.

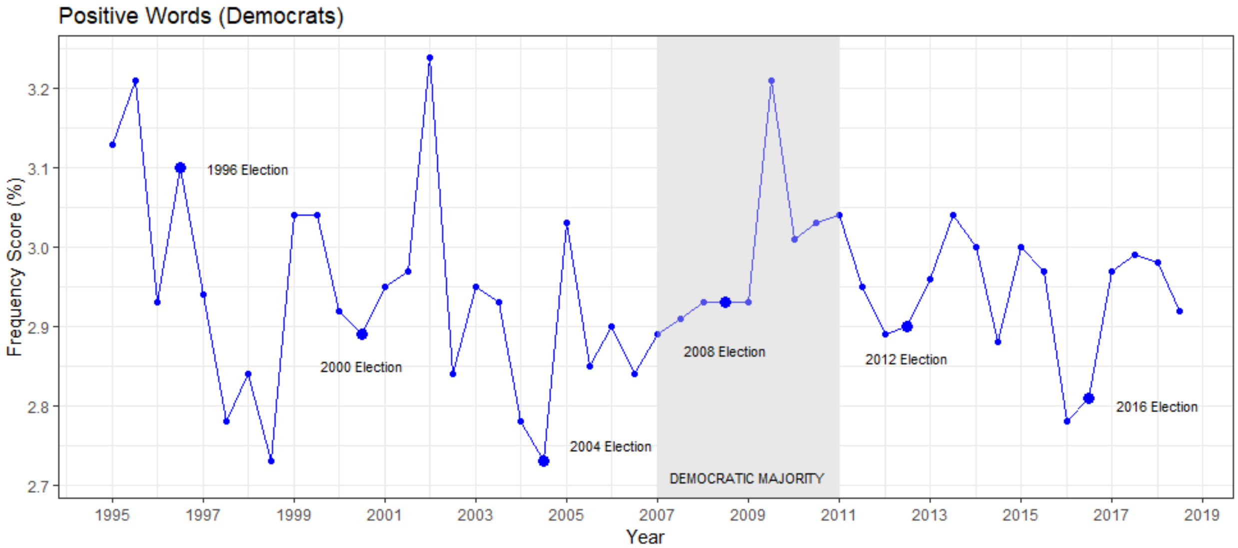

Figure 18: Positive Words (D) with Party Control and Presidential Elections

Just like the top one hundred words, Democratic control of the House from 2007 to 2011 sees little substantial change in word distribution. This is especially apparent in the positive Democratic words. The period between 2007 and 2011 is the most consistent and unchanging of the entire data frame. Excluding a strong spike in late 2009, each six-month selection has very little change in word frequency. Every other section of the data frame is characterized by strong surges and sharp drops, but it is only when the Democrats regain control of the House that the frequency changes begin to remain more consistent. Despite being more consistent, the word frequency during Democratic control is not much higher than when they were the minority party. This suggests that when Democrats are in the majority party, they are less erratic in their word choice than when they are out of power.

Negative Words

The final metric of word frequency is the negative Democratic words. When placed onto a line, the metric reveals a trend line demonstrated by *Figure 19.*

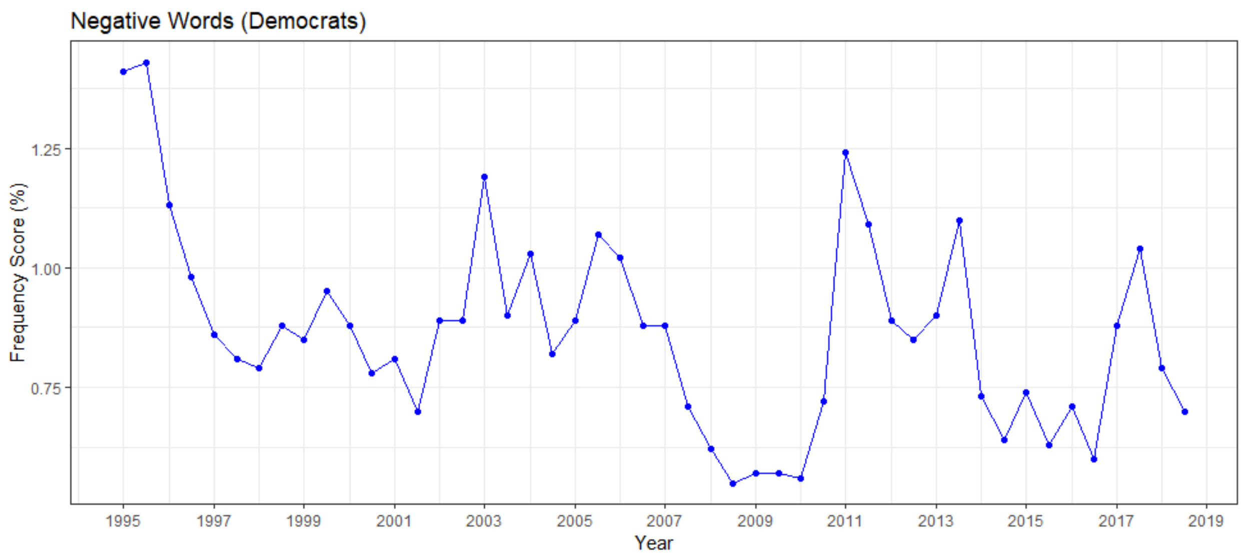

Figure 19: Negative Democratic Words

Just like the top one hundred words and the positive words, this trend begins with an extremely high rate of word frequency that begins to fall immediately after 1995. The trend then oscillates for several years until in 2005 when it starts a downward trend. This trend continues until 2008, where it levels off and remains consistent. After a brief calm the trend once again begins to spike up and down erratically from year to year.

The negative words metric does not match well with the trend of DW-NOMINATE polarization. This metric shows extreme variance and no overall consistent trend. The highest point of word frequency is in 1995 when polarization was comparatively low. The low period in word frequency from 2007 to 2010 is also not reflected in the DW-NOMINATE analysis. Rather that era is shown to be a time of rising polarization. The negative Democratic word frequency also ends at a level that is far below the average of the data frame, however if it matched with DW-NOMINATE polarization then the present day should be the highest point of word frequency. This metric offers little support to the notion that negative word frequency is correlated the median difference between the parties' DW-NOMINATE scores.

The PPI comparison yields similar results. The PPI and the negative word metric appear to be almost the opposite in certain circumstances. For example, in 2004 the PPI shows a polarization spike with either side rising towards a single peak. The negative words show the same year as a valley where either side descend towards the nadir in 2004. The negative words and the PPI also match in terms of the sharp rise in polarization around 2010 to 2011. However, the sharp drop in polarization around 2013 is not mirrored by Democratic negative word frequency, the chart actually demonstrates the opposite. The negative words show that 2013 was actually a strong surge in word frequency despite lowering levels of polarization. While post 2013 extends beyond the data frame of the PPI, the next few years represent a low point in word frequency. Overall, the negative Democratic word chosen for this study demonstrate little correlation with polarization defined by the PPI.

Comparing the negative word to party control and elections shows much stronger trends and relationships, as shown by *Figure 20.*

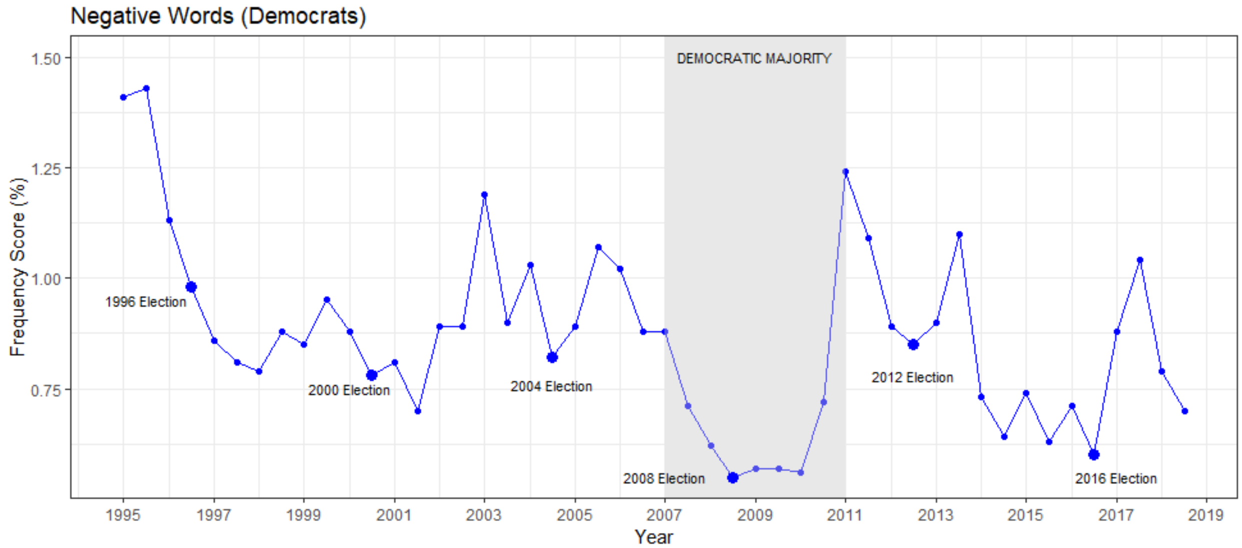

Figure 20: Negative Words (D) with Party Control and Presidential Elections

The most noticeable trend is the substantial decline in negative word during the Democratic control of the House from 2007 to 2011. In a short span of time after 2007, the frequency of negative word drops from over .8 percent to around .5 percent. The chart reaches its lowest and most consistent era when the Democrats have control of the House. Both the negative words and the positive word metric support the hypothesis that when Democrats are the majority part, there will be less variance from year to year in their word choice. While this does not strongly indicate more consistent word choice, it does suggest a higher amount of party discipline when Democrats are in control than when they are the minority party. Oddly enough, this is the inverse of the Republican trends. When the Republicans were out of power, there was a stark increase in word frequency. It is also important to note that the Democrat's majority does not appear to influence the amount of word frequency at any one point as much as the consistency of change over time. With the top one hundred and positive words, the Democratic majority was not associated with an increase or decrease in frequency, but rather with a flat trend line. The negative words metric is similar but also accompanied with a large drop in frequency score. This combined with the consistency of Republican word frequency during eras of Democratic majority support the notion that word choice and frequency are correlated with whether a party is in control of the House.

The presidential elections show a similar trend to the other word frequency metrics. With the exception of the 1996 election, each election is at the bottom of a ‘v’ formation, just like with the positive words. The most substantial post-election rise occurs after the 2016 election, where negative word frequency surged over 1%. The election that saw the smallest change in word frequency was the 2008 election, which occurred while Democrats were in control of the House. Looking at both the positive and negative word frequency, after an election there is almost always an ephemeral spike in word frequency. While the session immediately after the election is characterized a higher frequency score, this rise is short lived and after a year or two returns back to pre-election levels.

Chapter 4: Conclusion

Looking at the results of each metric, there is little support for the hypothesis that trends in polarization are closely tied to trends in the distribution of common words in congressional speeches. While some word frequency trends were reflected in polarization trends, just as many were dissonant. There is not sufficient evidence to reject the null hypothesis of this study. Despite this, the analyses of this research have demonstrated a much stronger correlation between word frequency and party control of the House. This trend is particularly pronounced with the Republican party, where the distribution of the top one hundred words became more clustered during Democratic control. During the Democratic majority, Republican word choice additionally contained higher frequencies of negative words and lower frequencies of positive words. These trends show a strong correlation between the political power shifts and word choice in Republican floor speeches.

The Democrats did not show as strong of a correlation with party control as Republicans, but instead showed a closer correlation with presidential elections. In almost every election cycle, across all metrics of Democratic word frequency, presidential elections were marked by a ‘v’ formation were word frequency scores dropped before an election and spiked immediately afterward. The Democrats also showed less change over time when they were in control of the House. This was demonstrated when the trends of each metric except positive words remained flat from 2007 to 2010. The strongest Democratic correlation occurred in the negative meat words metric, where the key negative words were far less frequent when Democrats were the majority party in the House. This demonstrates that the negative words chosen for this study are far less common when Democrats are in control of the House.

While the speech analysis did not strongly support the initial hypothesis of this study, it did yield other significant observations. Shifts in party majority corresponded with noticeable shifts across every Republican word frequency metric. The top one hundred words saw a consistent increase in word frequency score during the time that they were out of power. The positive Republican words also saw a sharp decrease in frequency at the same time. This decrease in the positive words was mirrored by an increase in negative Republican words. Among Democrats, the negative words analysis also saw a substantial drop in frequency score during Democratic control that rose sharply immediately after the 2011 election. The other Democratic metrics were less associated with increases or decreases in word frequency and more associated with placating frequency scores. Instead of increasing in sharp rigid peaks, the era where Democrats were in power is characterized by more gradual change from year to year. With both positive and negative words having nearly flat trend lines for the entire time that Democrats were the majority party in the House. This could suggest that Democrats are more varied with their speechmaking when they are the minority party and have less discordance with their rhetoric when they are in power.

Another interesting observation in the data gathered for this study is the trend of word frequency decreasing just before an election before sharply rising afterward. Both of these results provide avenues for further research. For one, less word frequency prior to an election could be the result of representatives taking more time to focus on their own elections rather than the party platform. This could be measured by an increase in constituent communications and speeches made on matters of local relevance. Each speech that relates to an individual representative's district will have its own unique concerns and words that are associated with local issues. If more representatives are making more varied speeches, that will result in less similarities among speeches and thus less word frequency. The Democrats also showed stronger reactions to elections than Republicans. The behavior of Democratic word frequency surrounding elections was a stronger reaction than with the Republicans. This suggests that the Democratic Party's speechmaking has more correlation with changes the Executive branch than the Republican's, whose rhetoric is more closely correlated with changes in the party that controls the House.

The results of the analysis indicate that, taken in totality, the higher consistency of word choice among party members is not strongly correlated to increased polarization. However, the metrics did reveal other observations that can lay the groundwork for further research. The most immediate path of future research would be to extend the analysis of DW-NOMINATE median difference from chapter two farther back than 1995. This could help paint a broader picture of Congressional polarization determine if the polarization of past Congresses is comparable to the modern era. One potential data frame would be to extend measurements of polarization back to the beginning of the post war period. Extending the comparison between Democrats and Republicans as a metric of polarization would be problematic if it is done before the mid-19th Century, since the Republican Party was not one of the two major parties. This study was intended to examine the polarization of the current two-party paradigm. As such, it is necessary to focus analysis on the eras of politics where Democrats and Republicans are the two major parties. Starting from the turn of the twentieth century, the DW-NOMINATE analysis shows a falling and increasing trend demonstrated by *Figure 21*.

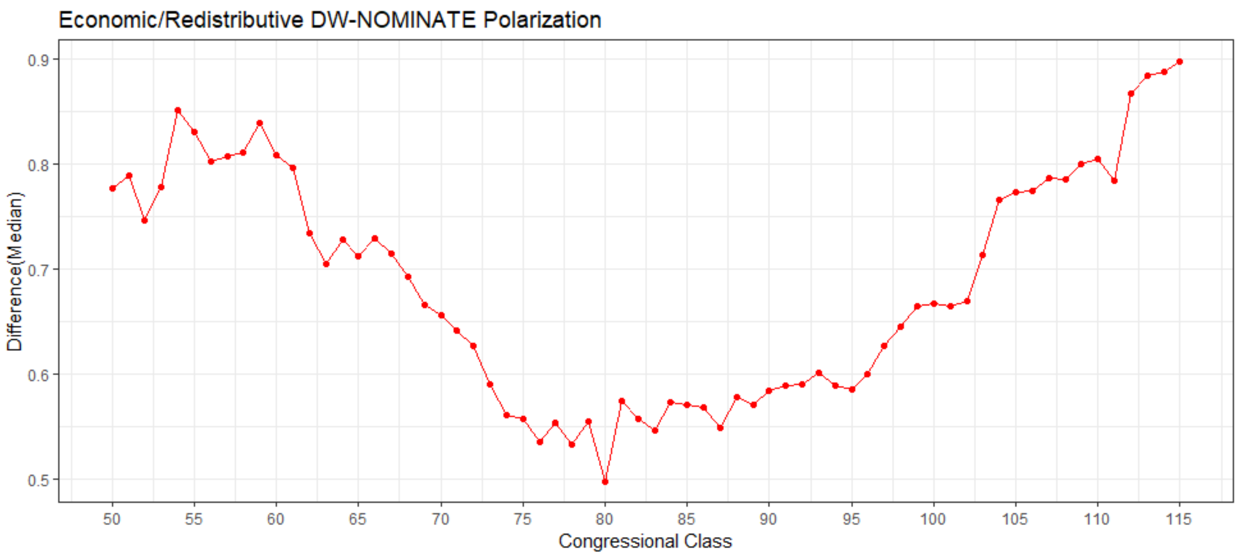

Figure 21: Economic DW-NOMINATE Polarization Since 1900

Much of the literature on polarization identified the phenomenon as a longstanding trend that has been continuous since the end of World War Two. This measure of polarization supports this notion. The low point of the chart occurs in the 80th Congress, which took office around 1947. This analysis also supports the idea that while current levels of polarization are comparatively high, they are not unprecedented. The trend shows that the early 1900s had similar levels of polarization as the current Congress. This observation is also supported by the social/racial dimension demonstrated in *Figure 22.*

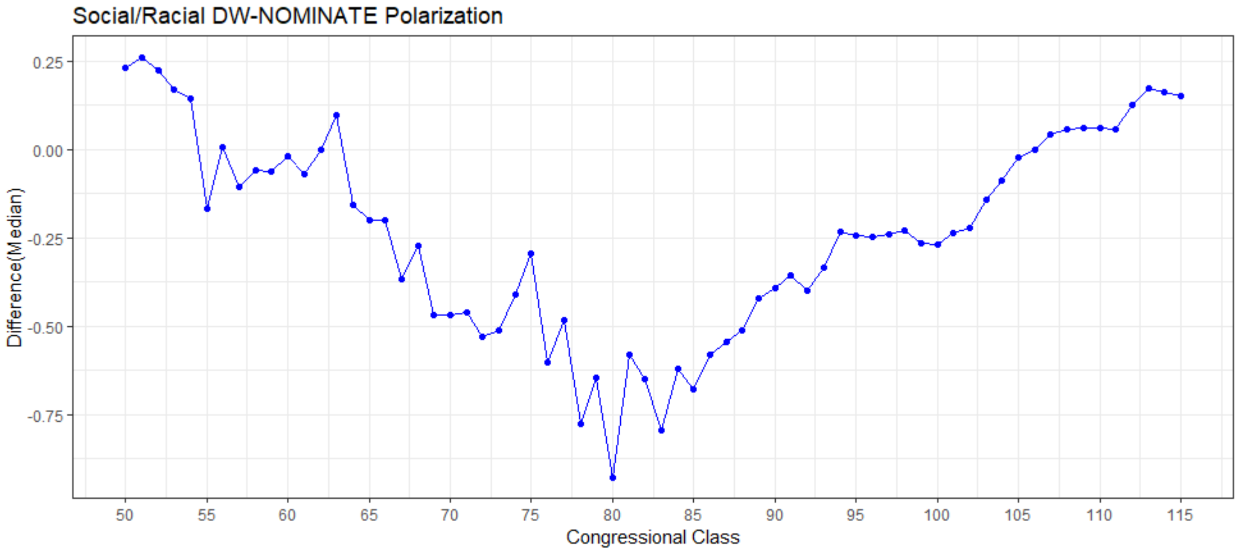

Figure 22: Social DW-NOMINATE Polarization Since 1900

Unlike the economic dimension of DW-NOMINATE, the social/racial dimension does not identify the current congressional class as the most polarized Congress. The congressional classes immediately after 1900 have higher or equivalent levels of social/racial polarization than the current Congress. Additionally, the social/racial difference becomes far less erratic in the post-war period. The pre-WWII era has dramatic shifts in the median difference between the two parties in each Congressional class. But after 1947, the trend smooths out and starts a steady upward climb until today. An option for future study would be to compare the word distribution of Congressional speechmaking during the low point of polarization in 1947 and the high points in 1900 and 2018. This would test the difference between congressional speechmaking during eras of extreme division and extreme unity.

One potential limitation with the construction of this study is that it only focuses on one aspect of one chamber. Polarization may manifest itself differently in other areas of Congress. To address this, one could potentially add press releases of Congresspersons to analyze how legislators communicate directly with their constituents. Furthermore, a supplemental study could be made on the speeches made by Senators, controlling for length of speech using representative sampling methods. These analyses could also be combined to evaluate how each body of Congress interacts with other, considering each chamber not as an independent study, but to see whether rises and falls in polarization in the House are tied to similar trends in the Senate.

A key aspect of this study is using specific words to capture the essence of party messaging, with research showing that certain words trend towards one party over the other -- the basis for the “meat words” analysis of word frequency. While gathering data I became curious whether the words chosen were successful at capturing the rhetorical essence of either party. In order for this to be the case, it would be necessary to show that the words identified as partisan words are more common among one party than the other. To investigate this, I ran the same analysis except I cross-referenced the results with the other party using the same words. For the Democrats I calculated the frequency score and then ran the same analysis on the word database for the Republican using the Democratic words as a filter. This metric only looks at a small section of the data frame, but still revealed some consistent trends that point to the effectiveness of the meat words metric.

Starting with the Democratic meat words, the words chosen for this analysis were successful in capturing words that were more popular among Democrats than Republicans, as demonstrated by *Table 1:*

	Positive Democratic Meat Words		Negative Democratic Meat Words	
	D	R	D	R
2013	**2.96**	2.79	**0.9**	0.68
	3.04	2.77	**1.1**	0.79
2014	**3**	2.76	**0.73**	0.68
	2.88	2.78	**0.64**	0.53
2015	**3**	2.84	**0.74**	0.57
	2.97	2.73	**0.63**	0.49
2016	**2.78**	2.79	**0.71**	0.53
	2.81	2.67	**0.6**	0.45
2017	**2.97**	2.85	**0.88**	0.57
	2.99	2.92	**1.04**	0.61
2018	**2.98**	2.86	**0.79**	0.54
	2.92	2.74	**0.7**	0.49

Table 1: Democratic Meat Word Differences

This table shows the frequency scores of the meat words since 2013. Each year has two data points, one for the first half and another for the second year. The left side shows the frequency of the Democratic meat words among Democrats, the right sides shows the frequency of the same words among Republicans. The positive meat words were more popular among Democrats than Republicans, even though the differences were not substantial. The only exception to this trend was in the first half of 2016, were the Democratic positive words were 0.01% higher among Republicans. In all other years, the difference in frequency score with the same set of words ranges from around 0.1% to 0.3%. The negative words also show a substantially higher popularity among Democrats, with the differences ranging from 0.3% to over 0.4%

Among Republicans, the words chosen for this analysis were less successful at capturing partisan rhetoric, as shown by *Table 2:*

	Republican Positive Meat Words		Republican Negative Meat Words	
	R	D	R	D
2013	**2.75**	2.93	**0.84**	0.61
	2.48	2.86	**0.7**	0.63
2014	**2.71**	2.81	**0.67**	0.54
	2.69	2.82	**0.64**	0.53
2015	**2.75**	2.75	**0.69**	0.59
	2.61	2.77	**0.54**	0.46
2016	**2.67**	2.76	**0.54**	0.52
	2.76	2.76	**0.49**	0.46
2017	**2.73**	2.65	**0.57**	0.62
	2.74	2.8	**0.82**	1.08
2018	**2.9**	2.95	**0.64**	0.64
	2.8	2.83	**0.51**	0.55

Table 2: Republican Meat Word Differences

The positive words chosen for this study were actually more common among Democrats by slim margins when taken as a percentage of the whole speech database. In almost every year except for 2017, the frequency scores were tied or higher among Democrats. With the negative words, they were more common among Republicans initially, but starting in 2017 they flipped and became more popular among Democrats by a small percentage. These results show that some of the words chosen for this study are not as unique to each party as initially suspected.

While some words were more common among the party that they were not intended to capture, this does not nullify the validity of this study's results. This analysis is primarily about capturing changes in word frequency within a party and less about what words are specific to each party. While in some years Republican meat words were more common among Democrats, this is not beholden on whether the words are more frequent from year to year. Asking what words are most common among each party and asking how word frequency changes from year to year are related but separable questions and this study primarily focuses on the latter.

The meat words analysis also brings up a host of other questions. For instance, since the analysis was conducted on a small body of words, the increase in the usage of one particular word may be responsible for the increase in the overall frequency score. I was curious to know whether the increases in frequency score were due an increase in usage of all the words across the board, the substantial increase in one particular word, or increases in words that relate to specific issue areas. To observe this, I chose several sequences on a chart that have an extreme decrease or increase in frequency and re-ran the analysis on the high point and the low point. This time, the analysis was more focused on the word counts themselves rather than the percentage that those words make of the whole body of text. The first two datapoints chosen were from the Republican positive word metric. From late 1996 to early 1997 the frequency score jumped from 2.53% to 3%. For this test, the words chosen were the top 17 words with counts over one hundred words. The difference and percent increase in word count is demonstrated by *Table 3.*

Word	Count 1996	Count 1997	Difference	Increase (%)
Freedom	455	1237	782	172
Family	1076	2769	1693	157
Businesses	425	1050	625	147
Debate	746	1700	954	128
Lives	534	1199	665	125
Rights	864	1797	933	108
Business	1068	2125	1057	99
Life	1016	1997	981	97
Families	1295	2458	1163	90
Power	541	997	456	84
Right	1446	2618	1172	81
Opportunity	964	1723	759	79
Children	2043	3569	1526	75
Help	1462	2126	664	45
Work	3132	4295	1163	37
Control	807	1097	290	36
Child	924	1198	274	30

Table 3: Positive Republican Word Increase 1996-1997

The table is organized in descending order based on the percentage increase from 1996 to 1997. The far-right column shows the percentage change from one year to the next. Some of the words saw increases ranging from thirty percent to over one hundred and seventy percent. The analysis showed that the increase in word frequency was not caused by the usage of one particular word, but rather by an increase of every word used for this metric. By far the most used word in the positive word analysis was “work,” which saw an increase from three thousand to over four thousand. Despite being the most used word, the increase in the usage of “work” was only around a 40 percent increase. The highest percent increase came from the words “freedom” and “family” which both saw an increase of over 150% from one session to the next.

An important factor of this experiment that separates it from the two points taken from the positive word metric is that both of these datapoints rest on roughly the same amount of speech data. One other potential factor that could be causing difference from year to year could be the size of the database itself. Some years have more speechmaking than others, and this may have some effect on the word counts and methods of speechmaking. To test this, the same analysis was repeated with two databases of roughly the same size. The two datapoints were taken from the negative words of Republicans, where the metric saw a decrease from 1.02 in the first half of 2001 to 0.54 in the second of half of 2001. This decrease also occurred in a similarly sized database, meaning the results cannot be explained by just the presence of more speeches. The data from 2001 demonstrated by *Table 4* shows a combination of words increasing and decreasing in usage.

Word	January-June	July-December	Difference	Change (%)
Destroy	105	259	154	147
Illegal	236	393	157	67
Consequences	170	225	55	32
Criminal	304	390	86	28
Threatened	102	121	19	19
Failure	198	229	31	16
Crisis	493	537	44	9
Democrat	231	228	-3	-1
Spent	625	592	-33	-5
Mandate	139	127	-12	-9
Democrats	388	309	-79	-20
Liberal	98	73	-25	-26
Bureaucracy	142	102	-40	-28
Imposed	122	79	-43	-35
Spend	778	491	-287	-37
Spending	1117	687	-430	-38
Taxes	1509	467	-1042	-69
Debts	145	44	-101	-70
Tax	6388	1548	-4840	-76
Taxed	138	21	-117	-85

Table 4: Negative Republican Word Decrease in 2001

As illustrated by the right-hand column of *Table 4,* some words saw increases in the second half of 2001, but most decreases substantially. For the negative words decrease during the 2001 session, the word frequency as a whole went down, meaning that the meat words made up a smaller portion of the total body of text. But words like “destroy” and “illegal” still saw substantial rises in the amount of times they were used. Many more words saw substantial decreases in use, such as “debts,” and “taxed.” In fact, every term relating to tax policy saw a substantial decrease in the second half of the year. This could be explained by the fact that budgetary discussions happen in the first few months of the year, which would make taxes and spending a more prominent topic. (GovInfo n.d.)

One key unexplored aspect of this analysis was the influence of the party message boards outlined in Chapter 2. Each party has subcommittees that are responsible for determining how the party platform will be expressed. These committees determine which words and frameworks will express the party's interests. (Harris 2005, 127) Another point of analysis to add to further research could be to investigate whether lawmakers follow the suggestions of these committees more during eras of high polarization. By comparing the suggestions of the committees to the words used by each party member on the floor, I could observe which party is closer in line with the party platform. Additionally, interviews with former and sitting Congresspersons could elucidate how much pressure they personally felt to conform to party messaging.

The presence of specific committees whose focus is on party messaging is also a very recent phenomenon. The entire scope of this study falls after the introduction of these committees. Another avenue of future research would be to replicate the analysis of this study but include speeches that happened before the introduction of party message boards. The presence of such messaging committees could have a measurable impact on the frequency which certain words are used on the floor and the overall consistency of congressional speechmaking. A before and after comparison of the Congressional speechmaking before and afterward could answer the question of whether the introduction of party message boards has an influence on the similarity between Congressional speeches.

The final dimension for future research regarding this study is the issue of how polarization effects the individual legislator. Many political actors hold their positions for multiple decades, meaning that they would be present for eras of high and low polarization. The question then becomes whether the distribution of words in their speeches becomes more skewed during times of high polarization or great controversy. The construction of this study lends itself to this future research well since the *Congressional Record* parser organizes each speech by the name of the speaker, making it easy to separate by individual.

The topics of this thesis relate to some of the most substantial debates in modern political discourse. It is common to lament the increase in polarization that maligns current government preventing compromise and encouraging hyper partisanship. But the very statement 'Congress is more polarized' carries with it a myriad of key definitional challenges. For one, that contention assumes that there is a way to accurately measure ideology, for one cannot say that the two parties have moved farther from each other without having some metric to measure ideological distance. Furthermore, it assumes that such a metric can be reliably extended to different eras of political time, in order to offer a fair comparison between one era and the next. From that point the question moves on to which issues are key to defining polarization as well as which issues can be separated from one another and which can be grouped together. For example, a concept like welfare or Medicaid could be considered a social issue or an economic one since it contains many elements of both.

Many scholars, using DW-NOMINATE among other metrics, have created methods of compartmentalizing and quantifying polarization. It is hard to say that any results gleaned from such analysis have actually captured the full scope of the issue. Over the course of this study, my most substantial takeaway was the difficulty in studying, measuring, and evaluating polarization. It is a multifaceted problem that displays its complexity through many subtle behaviors.

The issues arising from polarization are as much ideological as they are institutional: the behavior that voters expect from their representatives places them into a double bind with political failure on one end and lack of progress on the other. As discussed in the introduction to this study, the voters of both parties prefer obstinance over compromise. Each side is waiting for the other to blink, for doing so is a sign of weakness and lack of resolve. Conceding any demands of the opposite party is seen as a betrayal of a representative's beliefs. The result is a stalemate between two equally powerful parties where not making progress is in the immediate interests of both sides. There is less agency to resolve issues when representatives know that even if they do not find a compromise that their constituents will still support them for sticking to their beliefs. If the quality of political discourse is to be preserved, it will require a shift in the thinking surrounding politics. Compromise needs to be destigmatized, pride needs to be superseded by pragmatism. It is easy to be steadfast and inflexible, but much harder to be rational and understanding. Overcoming the ailments of polarization will require a commitment to negotiation and an understanding that no one can get everything they want.

Appendices

Appendix I: Democrat Frequency Scores

This table and *Appendix II* represent the data values that I collected to create the charts in Chapter III. The data contains the percent frequency scores for each metric. The first column represents what percent of the six month segment is concentrated top one hundred most frequent words. The second column is the same, except for the top fifty words. The third and fourth column are the frequency scores for the meat word metrics. Just like the top one hundred and top fifty, the value in each cell represents the percentage of the whole for each six month segment.

	Democrats			
	top100	top50	MW+	MW-
1995	20.23	13.25	3.13	1.41
	20.31	13.44	3.21	1.43
1996	19.73	12.86	2.93	1.13
	19.46	12.72	3.1	0.98
1997	19.2	12.59	2.94	0.86
	19.36	12.74	2.78	0.81
1998	18.76	12.12	2.84	0.79
	18.91	12.18	2.73	0.88
1999	19.02	12.23	3.04	0.85
	19.15	12.26	3.04	0.95
2000	19.16	12.27	2.92	0.88
	19.16	12.37	2.89	0.78
2001	19.67	12.68	2.95	0.81
	19.24	12.41	2.97	0.7
2002	19	12.12	3.24	0.89
	19.11	12.54	2.84	0.89
2003	19.61	12.43	2.95	1.19
	19.58	12.45	2.93	0.9
2004	19.43	12.44	2.78	1.03
	18.78	12.15	2.73	0.82

2005	18.53	11.84	3.03	0.89
	18.26	11.79	2.85	1.07
2006	18.73	12	2.9	1.02
	18.71	12.32	2.84	0.88
2007	19.8	12.77	2.89	0.88
	19.57	12.98	2.91	0.71
2008	19.32	12.55	2.93	0.62
	18.5	12.1	2.93	0.55
2009	19.12	12.33	2.93	0.57
	19.72	12.9	3.21	0.57
2010	19.5	12.75	3.01	0.56
	19.67	12.83	3.03	0.72
2011	19.83	12.54	3.04	1.24
	19.3	12.43	2.95	1.09
2012	19	12.2	2.89	0.89
	18.71	12.05	2.9	0.85
2013	19.18	12.28	2.96	0.9
	19.71	12.88	3.04	1.1
2014	18.88	12.19	3	0.73
	18.57	12.12	2.88	0.64
2015	18.67	12	3	0.74
	18.37	11.82	2.97	0.63
2016	18.2	11.81	2.78	0.71
	18.49	12	2.81	0.6
2017	19.71	12.44	2.97	0.88
	19.29	12.56	2.99	1.04
2018	18.5	12.05	2.98	0.79
	18.6	12.11	2.92	0.7

Appendix II: Republican Frequency Scores

	Republicans			
	top100	top50	MW+	MW-
1995	20.45	13.66	2.59	1.28
	20.51	13.75	2.38	0.99
1996	19.64	13.13	2.54	1.13
	18.95	12.59	2.53	0.75
1997	19.95	13.23	3	1.14
	20.04	13.36	2.66	1.05
1998	19.16	12.7	2.68	1.04
	19.1	12.57	2.54	0.86
1999	18.91	12.34	2.51	0.92
	19.39	12.72	2.52	0.92
2000	19.51	12.71	2.61	1.02
	19.31	12.66	2.56	0.87
2001	19.73	12.85	2.91	1.02
	19.41	12.81	2.71	0.54
2002	19.09	12.46	2.78	0.82
	18.91	12.43	2.48	0.56
2003	19.03	12.34	2.91	0.82
	19	12.13	2.66	0.66
2004	19.21	12.36	2.75	0.88
	18.62	12.25	2.57	0.62
2005	18.59	12.06	2.64	0.68
	18.27	11.99	2.68	0.68
2006	18.88	12.33	2.63	0.79
	18.65	12.29	2.62	0.55
2007	19.94	12.89	2.48	1.07
	19.93	13.11	2.52	0.86
2008	19.61	12.8	2.49	0.81
	19.13	12.52	2.31	0.61
2009	19.86	12.62	2.4	1.07
	20.43	13.26	2.38	0.94
2010	20.59	13.7	2.5	0.98
	19.82	13	2.55	1.06
2011	20.53	13.29	2.7	0.99
	20.54	13.4	2.6	1
2012	19.6	12.7	2.6	0.85
	19.66	12.69	2.71	0.96

2013	19.67	12.76	2.75	0.84
	18.61	12.3	2.48	0.7
2014	18.84	12.4	2.71	0.67
	19.15	12.72	2.69	0.64
2015	18.92	12.53	2.75	0.69
	18.62	12.16	2.61	0.54
2016	18.68	12.23	2.67	0.54
	18.96	12.57	2.76	0.49
2017	18.76	12.4	2.73	0.57
	19.07	12.5	2.74	0.82
2018	18.89	12.39	2.9	0.64
	18.87	12.52	2.8	0.51

<u>Appendix III: Top 100 Republican Words: July-December 2018</u>

Word	Count						
Years	3987	Legislation	1982	Served	1309	Good	1024
Today	3669	Colleagues	1904	Need	1302	Since	1021
People	3520	Family	1852	Security	1298	Congressional	1014
President	3065	Life	1654	Health	1277	California	1010
Support	2973	Great	1638	Get	1242	Business	999
Work	2973	Members	1636	Balance	1233	Even	998
States	2890	Honor	1627	Senate	1228	Children	984
Community	2855	Important	1590	Leadership	1218	Services	984
Service	2808	Texas	1537	Office	1217	Better	968
Many	2683	Back	1533	Every	1180	Families	968
Act	2672	Member	1525	Home	1166	Local	960
State	2595	District	1523	Department	1139	Think	957
United	2437	Government	1520	Young	1139	Right	949
Committee	2399	County	1513	Day	1132	Bipartisan	938
Thank	2379	Well	1505	Last	1127	Join	937
Rise	2355	School	1496	Nation	1111	Working	934
House	2333	Ask	1474	America	1106	Judge	932
New	2331	Program	1461	Two	1076	High	930
Year	2259	Help	1449	Continue	1068	Provide	930
American	2096	Know	1440	Americans	1054	Pro	928
Country	2048	Federal	1406	Tax	1052	Washington	922
Yield	2020	Recognize	1392	Urge	1047	Military	912
National	2011	Law	1386	Across	1045	Veterans	910
Like	2002	World	1377	War	1034	Small	906
First	1982	Public	1350	Including	1029	Worked	897

Appendix IV: Top 100 Democratic Words: July-December 2018

Word	Count						
People	3212	Public	1606	Senate	1082	America	895
President	3044	Like	1574	Great	1081	Education	892
Years	2886	Legislation	1556	Get	1057	Working	892
Community	2599	Important	1404	County	1056	Rights	881
Support	2543	Member	1400	Trump	1048	Don’t	877
Today	2460	Health	1379	World	1046	Ask	874
American	2345	Know	1335	Balance	1042	Bipartisan	872
Work	2334	Federal	1321	City	1039	Programs	867
Act	2274	Need	1309	Every	1029	Two	867
Many	2244	Government	1306	Even	1023	Administration	864
New	2200	Americans	1303	Judge	1022	Good	863
States	2136	Children	1294	District	1019	Million	845
Colleagues	1965	Life	1279	Tax	1010	Must	842
Service	1907	Back	1267	Leadership	1006	Republican	842
Yield	1857	Program	1252	Served	964	Justice	830
Rise	1793	Help	1236	Percent	962	Women	825
State	1791	Well	1218	Including	960	Local	820
Year	1774	Family	1213	Last	955	Continue	814
United	1753	Members	1196	California	946	Services	813
Committee	1749	Law	1192	Right	937	Across	810
National	1689	School	1192	Communities	924	Since	806
Country	1685	Vote	1170	Urge	920	Provide	796
Thank	1683	Security	1127	Court	909	Join	778
First	1646	Families	1109	Department	908	Nation	772
House	1619	Honor	1092	Day	905	Worked	766

References

Azzimonti, Marina. 2013. "The Political Polarization Index." *Federal Reserve Bank of Philadelphia*.

Box-Steffensmeier, Janet M., and David Canon. 2015. "Party Loyalty and the Potential Mechanisms of Party Discipline." In *Party Discipline in the U.S. House of Representatives*, ed. Kathryn Pearson. University of Michigan Press, 53–74. https://www.jstor.org/stable/10.3998/mpub.4402299.6.

Cameron, Charles, William Howell, Scott Alder, and Charles Riemann. 1997. *Divided Government and the Legislative Productivity of Congress, 1945-1994*. Washington DC: American Political Science Association.

Campbell, James. 2016. *Polarized: Making Sense of a Divided America*. Princeton University Press.

Carroll, Royce et al. 2009. "Measuring Bias and Uncertainty in DW-NOMINATE Ideal Point Estimates via the Parametric Bootstrap." *Political Analysis* 17(3): 261–75.

"Congressional Biographical Directory." http://bioguide.congress.gov/biosearch/biosearch1.asp.

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu, and Stefan Kaufman. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42(1): 31–55.

Dietrich, Bryce, Matthew Hayes, Saumil Dharia, and Stella Wancke. 2017. *Racial Rhetoric in Black and White: Symbolic and Substantive References in U.S. House Speeches*. http://www.brycejdietrich.com/files/working_papers/DietrichHayes_civil_rights.pdf.

Fiorina, Morris, Samuel J. Abrams, and Jeremy C. Pope. 2011. *Culture War? The Myth of Polarized America*. 3rd ed. New York: Longman.

Gingrich, Newt. 1996. "Newt Gingrich's 1996 GOPAC Memo."

GovInfo. "Budget of the United States Government." https://www.govinfo.gov/help/budget (February 18, 2019).

Graetz, Michael J., and Ian Shapiro. 2005. "A Political Mystery." In *Death by a Thousand Cuts*, The Fight over Taxing Inherited Wealth, Princeton University Press, 3–11. https://www.jstor.org/stable/j.ctt7t672.3 (February 27, 2019).

Graham, John D., ed. 2016. “Midterm Massacres.” In *Obama on the Home Front,* Domestic Policy Triumphs and Setbacks, Indiana University Press, 307–41. https://www.jstor.org/stable/j.ctt1ddr633.13 (November 19, 2018).

Harris, Douglas B. 2005. “Orchestrating Party Talk: A Party-Based View of One-Minute Speeches in the House of Representatives.” *Legislative Studies Quarterly* 30(1): 127–41.

———. 2013. *Lets Play Hardball: Politics to the Extreme*. eds. Scott Frisch and Sean Kelly. New York: Palgrave Macmillan.

Hunter, James D., Alan Wolfe, E.J. Dionne, and Michael Cromartie. 2006. *Is There a Culture War?: A Dialogue on Values and American Public Life*. Brookings Institution Press.

Jones, David R. 2001. “Party Polarization and Legislative Gridlock.” *Political Research Quarterly* 54(1): 125–41.

Kelly, Sean. 1993. *Divided We Govern? A Reassessment.* Polity 25.

Lenchner, Paul. 1976. “Congressional Party Unity and Executive-Legislative Relations.” *Social Science Quarterly* 57(3): 589–96.

Maltzman, Forrest, and Lee Sigelman. 1996. “The Politics of Talk: Unconstrained Floor Time in the U.S. House of Representatives.” *The Journal of Politics* 58(3): 819–30.

Mayhew, David R. 1991. *Divided We Govern*. New Haven: Yale University Press.

McCarthy, Nolan, Keith T. Poole, and Howard Rosenthal. 2016. *Polarized America: The Dance of Ideology and Unequal Riches*. 2nd ed. Cambridge, Massachusetts: The MIT Press.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16(4): 372–403.

Morris, Jonathan S. 2001. “Reexamining the Politics of Talk: Partisan Rhetoric in the 104th House.” *Legislative Studies Quarterly* 26(1): 101–21.

Nicholas, Judd, Jeremy Carbaugh, and Lindsay Young. 2017. *Congressional-Record: A Parser for the Congressional Record*. Chicago, IL.

Pennebaker, James W., Rodger J. Booth, and Martha E. Francis. 2007. *Linguistic Inquiry and Word Count: LIWC 2007*. Austin, Texas: Pennebaker Conglomorates inc. www.liwc.net.

Pew Research Center. 2014a. “Growing Ideological Consistency.” *Pew Research Center for the People and the Press*. http://www.people-press.org/2014/06/12/section-1-growing-ideological-consistency/ (February 20, 2018).

———. 2014b. “Political Polarization in the American Public.” http://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/ (February 17, 2019).

———. 2018. “The Tone of Political Debate, Compromise with Political Opponents.” http://www.people-press.org/2018/04/26/8-the-tone-of-political-debate-compromise-with-political-opponents/ (February 23, 2019).

Poole, Keith, and Howard Rosenthal. 2018. “DW-Nominate Scores.” https://voteview.com/congress/house.

Roberts, Jason M., and Steven S. Smith. 2003. “Procedural Contexts, Party Strategy, and Conditional Party Voting in the U.S. House of Representatives, 1971-2000.” *American Journal of Political Science* 47(2): 305–17.

Robinson, David, and Julia Silge. 2018. *Text Mining with R: A Tidy Approach*. O’Reilly. https://www.tidytextmining.com/tfidf.html.

Theriault, Sean M. 2008. *Party Polarization in Congress*. New York: Cambridge University Press.

Wood, Dan B., and Soren Jordan. 2017. *Party Polarization in America: The War Over Two Social Contracts*. New York: Cambridge University Press.

Yu, Bei. 2014. “Language and Gender in Congressional Speech.” *Literary and Linguistic Computing* 29(1): 118–32.